# A Large-Scale Test of Free-Energy Simulation Estimates of Protein-Ligand Binding Affinities.

Mikulskis, Paulius; Genheden, Samuel; Ryde, Ulf

*Published in:*
Journal of Chemical Information and Modeling

*DOI:*
10.1021/ci5004027

2014

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim



# A Large-Scale Test of Free-Energy Simulation Estimates of Protein–Ligand Binding Affinities

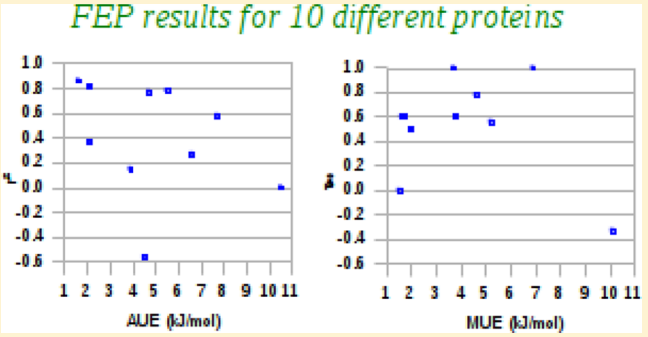
Paulius Mikulskis,[†] Samuel Genheden,[‡] and Ulf Ryde*[,†]

[†]Department of Theoretical Chemistry, Lund University, Chemical Centre, P.O. Box 124, SE-221 00 Lund, Sweden

[‡]Chemistry, University of Southampton, Highfield, SO17 1BJ, Southampton, United Kingdom

Supporting Information

**ABSTRACT:** We have performed a large-scale test of alchemical perturbation calculations with the Bennett acceptance-ratio (BAR) approach to estimate relative affinities for the binding of 107 ligands to 10 different proteins. Employing 20-Å truncated spherical systems and only one intermediate state in the perturbations, we obtain an error of less than 4 kJ/mol for 54% of the studied relative affinities and a precision of 0.5 kJ/mol on average. However, only four of the proteins gave acceptable errors, correlations, and rankings. The results could be improved by using nine intermediate states in the simulations or including the entire protein in the simulations using periodic boundary conditions. However, 27 of the calculated affinities still gave errors of more than 4 kJ/mol, and for three of the proteins the results were not satisfactory. This shows that the performance of BAR calculations depends on the target protein and that several transformations gave poor results owing to limitations in the molecular-mechanics force field or the restricted sampling possible within a reasonable simulation time. Still, the BAR results are better than docking calculations for most of the proteins.




## INTRODUCTION

One of the prime goals of computational chemistry is to develop accurate methods to estimate the affinity of a small molecule (L) binding to a biomacromolecule (R), i.e. the free energy of the reaction

$$L + R \rightarrow LR \quad (1)$$

Such a method would find great use in medicinal chemistry if the binding constant of a drug candidate could be accurately predicted without synthesizing it. It would typically then be enough to estimate relative affinities, i.e. to predict a correct ranking of a series of homologous ligands. On the other hand, the requirements on the accuracy is high—the method should be able to discriminate between drugs that have an affinity difference of about 4 kJ/mol, corresponding to a difference of a factor of 5 in the binding constant.

Alchemical perturbation methods, e.g. free-energy perturbation or thermodynamic integration, are computer protocols that combine a thermodynamic cycle with free-energy estimates of transforming one state into another, e.g., one ligand into another, by employing molecular dynamics or Monte Carlo simulations.[1] Alchemical methods are based on an exact statistical mechanical formalism and should in principle give a correct estimate of the binding free energy, provided that the energy function is correct and the sampling is perfect. However, extensive sampling is affordable only at the molecular-mechanics level, and even at that level, perfect sampling is not practically feasible. Therefore, alchemical methods also involve approximations and need to be carefully validated against experimental data sets to identify strengths and weaknesses of various approaches. Typically, alchemical methods are considered to be too costly for drug development, because they require a slow transformation of one state into another, through a series of unphysical, intermediate states,[2] although a few successful applications to drug development have been presented.[1,3] Therefore, cheaper alternatives such as scoring functions and end-point methods have been much more used for practical drug development. Although such approaches sometimes have given useful results, they are typically less accurate than alchemical methods.[4,5]

A natural approach to speed up molecular simulations is to consider only a restricted number of atoms around the site of interest. Such approaches have been used for a long time.[6−8] Recently, we have shown the efficiency and accuracy of such an approach for relative binding free-energies calculated by alchemical perturbation methods—all atoms outside a sphere of 20 Å of the ligand could be ignored without changing the calculated affinities by more than 1 kJ/mol.[9] Moreover, the efficiency can be further increased by a careful analysis of the simulation time and the required number of intermediate states.[10] However, the impact of these results is restricted because the approach was evaluated only on a single test case, viz., the binding of nine inhibitors to the blood clotting factor

**Received:** July 7, 2014

   dx.doi.org/10.1021/ci5004027 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

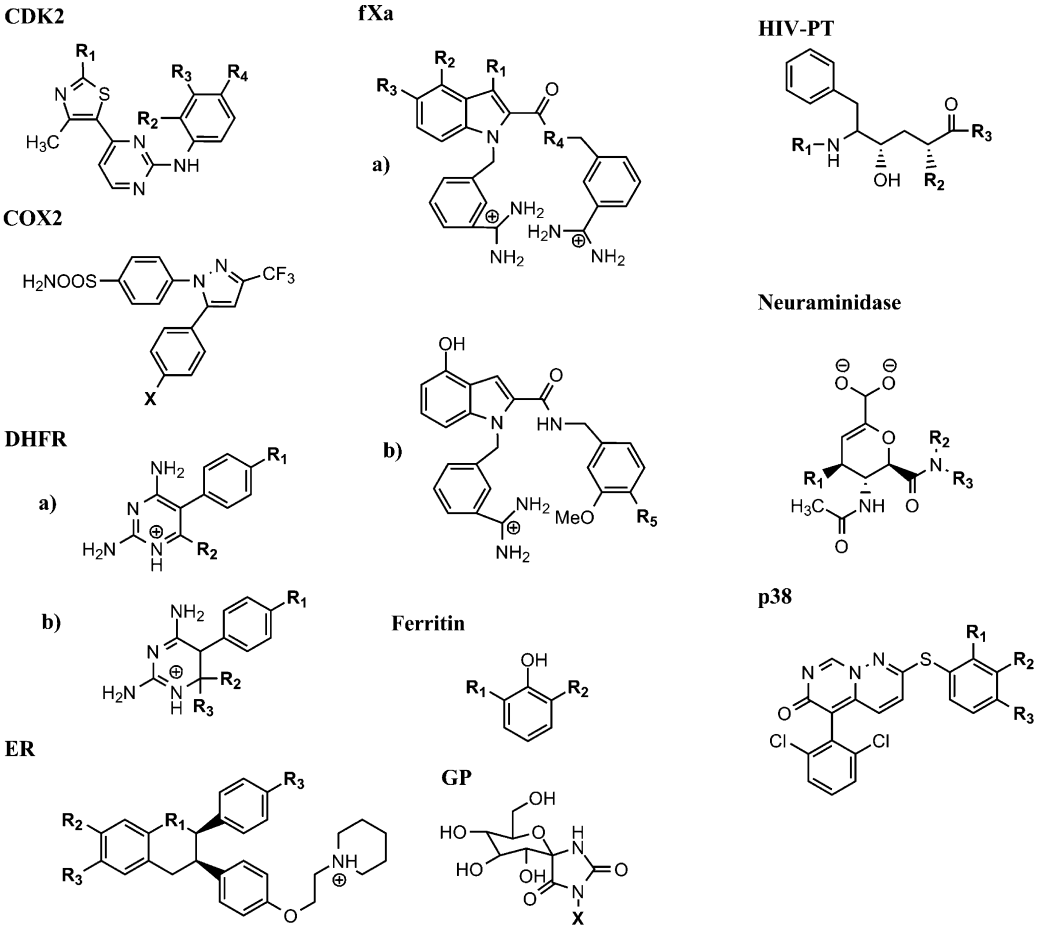



**Figure 1.** Ligand structures considered in this study. The sites considered for transformations are shown in bold face and are denoted $R_1$, $R_2$, etc., or X if it is a single site. For DHFR and fXa, there are two different scaffolds, denoted by a and b.

Xa (fXa). In this study, we therefore present a large-scale test of the protocol optimized for fXa and investigate whether the protocol is general. We have selected a diverse set of 107 ligands binding to 10 proteins, viz., cyclin-dependent kinase 2, cyclooxygenase-2, dihydrofolate reductase, estrogen receptor, factor Xa, ferritin, glycogen phosphorylase, human immunodeficiency virus protease, neuraminidase, and p38$\alpha$ MAP kinase. Our aim is to investigate whether a simple alchemical perturbation protocol for relative binding affinities is applicable to a large and diverse set of protein−ligand complexes, without any particular effort to address protein-specific issues or any extensive sampling, thereby aiming at a setting suitable for drug design. We also compare the results with those obtained by standard docking methods.

## METHODS

**System Preparation.** Ten proteins were considered in this study: cyclin-dependent kinase 2 (CDK2), cyclooxygenase-2 (COX2), dihydrofolate reductase (DHFR), estrogen receptor (ER), factor Xa (fXa), ferritin, glycogen phosphorylase (GP), human immunodeficiency virus protease (HIV-PT), neuraminidase (NA), and p38$\alpha$ MAP kinase (p38). In total, 107 ligands were studied, and template structures of these are shown in Figure 1. The systems were selected to obtain a wide range of proteins and ligands for which relative affinities can be estimated by alchemical perturbation methods.[11] In particular, we considered only charge-preserving transformations, and the perturbations were rather small, ranging from H → F to H → $CH_2CH_2CH_3$ or H → $CF_3$, i.e., introducing at most four non-hydrogen atoms. All except two involved perturbations at a single site. The proteins represent many different classes and display varying binding sites as shown in Figures S1 and S2. All ligands represent pharmaceutical compounds. Most of the systems have been studied with computational tools before,[12−18] and all inhibitors bind by noncovalent interactions.

The proteins were prepared in the following way: A representative crystal structure was selected for each protein (specified in Table S1), and they were protonated using the leap module of Amber11,[19] assuming a pH of 7 (i.e., Asp and Glu were negatively charged; Arg, Lys, and possibly His positively charged; and the other residues neutral). The protonation of the His residues was determined by investigating the hydrogen-bond network and solvent accessibility, and the assigned protonation states are listed in Table S1. No counterions were used to neutralize the systems. All proteins were described with the Amber99SB force field,[20] except fXa, which was described with the Amber99 force field,[21] following our earlier studies.[9,10] As the two force fields share the same charges and differ only in the protein backbone parameters, we expect very little difference between the results of the two force fields.

Ligands for which no crystal structure was available (see Table S1) were built by manually adding or removing atoms of ligands in available crystal structures, assuming a similar binding mode. The ligands were described with the general Amber force field,[22] and protons were added to the ligands using


dx.doi.org/10.1021/ci5004027 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

UCSF Chimera.[23] Charges were obtained using the restrained electrostatic potential method:[24] The ligands were optimized with the semiempirical AM1 method,[25] followed by a single-point calculation at the Hartree−Fock/6-31* level to obtain the electrostatic potentials, sampled with the Merz−Kollman scheme.[26] The potentials were then used by Antechamber[19] to calculate the charges.

Two setups of the proteins were employed. In the first,[9] the protein−ligand complexes or the free ligand were solvated in a sphere of TIP3P water molecules[27] with a radius of 20 Å, centered on the coordinate center of the ligand. Protein residues outside the sphere were kept in the simulations but were restrained to the starting coordinates with a force constant of 837 kJ/mol/Å$^2$, and they were excluded from the calculations of the nonbonded interactions.[9] This setup will be called spherical, and it was employed for all proteins.

In the second setup,[10] we employed instead periodic boundary conditions (therefore, this setup will be called periodical in the following), and the entire protein was included in the calculations. The protein−ligand complex or the free ligand was put into a truncated octahedral periodic box of TIP3P water molecules extending at least 10 Å from the solute.

**Free-Energy Calculations.** The relative binding free energy between two ligands, L1 and L2, $\Delta\Delta G = \Delta G_{bind}(L2) - \Delta G_{bind}(L1)$, was calculated for 91 pairs of ligands (see Table S2). The studied transformations were selected based on the availability of experimental data and computational convenience and thus do not involve all possible combinations of the selected 107 ligands. We employed a thermodynamic cycle that relates $\Delta\Delta G$ to the free energy of alchemically transforming L1 into L2 when they are either bound to the protein, $\Delta G_{bound}$, or free in solution, $\Delta G_{free}$,[28]

$$\Delta\Delta G = \Delta G_{bind}(L2) - \Delta G_{bind}(L1) = \Delta G_{bound} - \Delta G_{free} \quad (2)$$

$\Delta G_{bound}$ and $\Delta G_{free}$ were estimated by the Bennett acceptance-ratio method[29] (BAR) by dividing the transformation into a discrete number of states, described by a coupling parameter $\lambda$. Energies were also calculated by thermodynamic integration (TI) and exponential averaging (results described in the Supporting Information).

For the spherical setup, we followed our recent suggestion[9,10] to simulate only the end states ($\lambda$ = 0 or 1) and a single intermediate state at $\lambda$ = 0.5. The electrostatic and van der Waals interactions were transformed simultaneously in the simulation by using soft-core potentials for disappearing atoms,[30,31] as recently implemented in the Q programs.[9,32] The perturbations used a single-topology approach, transforming disappearing atoms into dummy atoms. An automatic script to set up the free-energy calculations can be found in http://www.teokem.lu.se/~ulf/Methods/fepq.html.

The periodic simulations employed the Amber11 software[19] and 13 intermediate states ($\lambda$ = 0.00, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, and 1.00). The electrostatic and van der Waals interactions were transformed simultaneously in the simulation by using soft-core potentials for disappearing atoms and a dual-topology approach.[33] An automatic script to also set up these calculations can be found in http://www.teokem.lu.se/~ulf/Methods/rel_free.html.

**Simulations with Spherical Systems.** The minimizations were performed with the sander module of Amber11,[19] and the MD simulations were carried out using Qdyn5 in the Q software suite.[32] The temperature was kept constant at 300 K using a weak-coupling thermostat.[34] In the minimization, a 20 Å cutoff was used for the nonbonded interactions, whereas in the MD simulations a 10 Å cutoff was used, except for interactions with the ligand, for which no cutoff was applied. In the MD simulations, long-range electrostatics were treated with the local reaction-field approach,[35] and water molecules were subjected to polarization and radial restraints as implemented in the Q software.[32] When simulating the protein−ligand complexes, solute atoms outside the simulation sphere were kept fixed at their initial positions using a strong harmonic restraint (837 kJ/mol/Å$^2$), and solute atoms in the outermost 2 Å shell were weakly restrained (84 kJ/mol/Å$^2$). When simulating the free ligand, the center of mass of the ligand was weakly restrained (22 kJ/mol/Å$^2$) to the center of the simulated sphere. The nonbonded pair list was updated every 25 steps. SHAKE[36] was applied to all bonds involving hydrogen atoms, and a time step of 2 fs was used.

The free-energy simulations were performed as follows: The system at $\lambda$ = 1 was minimized using 100 steps of steepest descent and then equilibrated, first using a 20 ps simulation in which all hydrogen atoms and water molecules were allowed to move, and the rest of the atoms were restrained toward their starting positions with a harmonic restraint of 105 kJ/mol/Å$^2$, and a 30 ps unrestrained simulation. Thereafter, the free-energy simulations were started at different $\lambda$ values. They consisted of 20 ps restrained equilibration, 500 ps (200 ps in case of fXa) unrestrained equilibration, and 1 ns production. Energy differences for BAR were sampled every 10 ps, because this is close to the estimated correlation time of these energies.

**Simulations with Periodic Systems.** The minimizations and MD simulations were performed with the sander module of Amber11,[19] except the calculations for the $\lambda$ = 0.00 and 1.00 states, which were performed with the pmemd module of a prerelease of Amber14.[37] The temperature was kept constant at 300 K using a Langevin thermostat[38] with a collision frequency of 2.0 ps$^{-1}$, and the pressure was kept constant at 1 atm using a weak-coupling isotropic algorithm[34] with a relaxation time of 1 ps. Particle-mesh Ewald summation[39] with a fourth-order B spline interpolation and a tolerance of 10$^{-5}$ was used to handle long-range electrostatics. The cutoff for nonbonded interactions was set to 8 Å, and the nonbonded pair list was updated every 50 fs. The SHAKE algorithm[36] was used to constrain bonds involving hydrogen atoms so that a 2 fs time step could be used.

The alchemical perturbation simulations were performed in the following way: The system at each $\lambda$ value was minimized for 100 cycles of steepest descent, with all atoms except water molecules and hydrogen atoms restrained to their start position with a force constant of 418 kJ/mol/Å$^2$. This was followed by a 50 ps NPT simulation, and a 500 ps NPT equilibration without any restraints. Finally, a 1 ns production simulation was run. Energy differences for BAR were sampled every 10 ps. In a few cases, sporadic SHAKE problems were encountered (related to the compiler[40]), which were solved by not constraining any bond lengths for the ligand and reducing the time step to 1 fs.

**Docking Calculations.** For the same systems, docking calculations were performed with two different software. First, we used GOLD version 5.1[41] and the ChemScore scoring function.[42,43] The binding site was defined as all residues within 10 Å of the ligand. The starting structures were the same as for MD, but all water molecules were stripped off. Ten repeats of 10 genetic algorithm runs were performed. The details of each run (population size, number of operations, etc.) were selected


dx.doi.org/10.1021/ci5004027 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

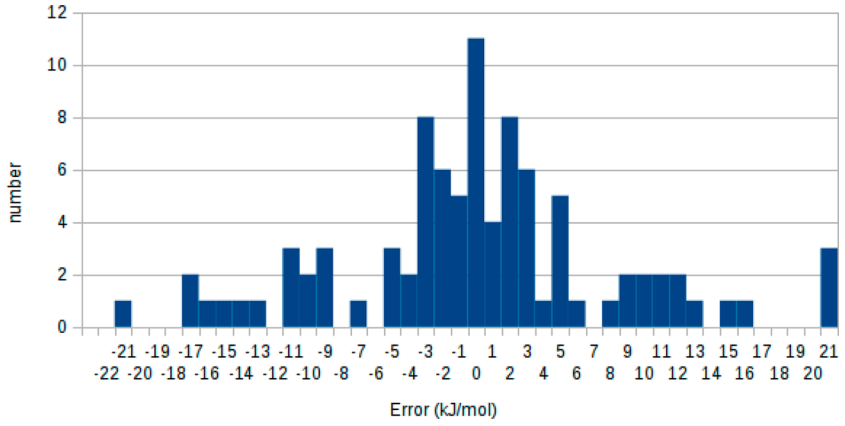



**Figure 2.** Histogram over the errors in the 91 relative binding free energies compared to experiments for the spherical setup using three $\lambda$ values.

automatically by GOLD. The highest score among the 10 repeats was used as the final prediction. If the root-mean-squared deviation (RMSD) with respect to the pose in the starting structure was higher than 2 Å, we also took the score of the pose that had the lowest RMSD among the top solutions of the 10 repeats. If the RMSD still was higher than 2 Å, we instead took the highest score of all solutions (not only the top one) that had an RMSD lower than 2 Å (or the lowest RMSD if no such solution was found).

The second set of docking calculations was performed with the DOCK 6.5 software.[44] The starting structures were the same as for the GOLD calculations. The binding site was defined to be within 10 Å of the ligand (except for HIV-PT, for which it was 13 Å, owing to the size of ligands). The ligand charges were calculated with antechamber using the AM1-BCC approach. Default Dock6 parameters were used throughout. The grid calculations used the bump filter with an overlap of 0.75. Reported scores were obtained by Grid score, which is a discretization of the Amber nonbonded interaction energies.[44]

**Uncertainties and Quality Metrics.** The uncertainties of the free-energy estimates were obtained by nonparametric bootstrap sampling (using 100 samples) of the work values in the BAR calculations.

The quality of the binding-affinity estimates compared to experimental data were quantified using the average unsigned error (AUE), median unsigned error (MUE), the correlation coefficient ($r^2$; a negative sign indicates that $r$ is negative), and Kendall's rank correlation coefficient ($\tau$) calculated only for the transformations explicitly simulated (i.e., not for all pairs of ligands that can be constructed from these values). The latter coefficient was also calculated after removing predicted and experimental relative affinities that were not significantly different from zero at the 90% level ($\tau_{90}$).[45] For COX2, ER, fXa, NA, and p38, no experimental uncertainties were reported, and we then assumed an experimental uncertainty of 1.7 kJ/mol.[46] The uncertainties of the quality metrics were obtained by a parametric bootstrap (500 samples) using the estimates and their uncertainties.

## RESULTS AND DISCUSSION

**Results with the Spherical Setup.** We have estimated the relative free energies between 91 pairs of ligands from 10 different proteins employing alchemical free-energy perturbations and the standard thermodynamic cycle with the two ligands either bound to the protein or free in solution.[28] Free-energy differences were calculated with the Bennett acceptance-ratio (BAR) method. The studied transformations range from small transformations, such as H → F, to relatively large transformations, like H → propyl, i.e., changes typical at the level of lead optimization. To start with, we used our recently suggested approach to minimize the computational effort using spherical systems and taking into account only protein atoms within 20 Å of a central atom of the ligand.[9,10] Moreover, we used only the two end points and a single intermediate state at $\lambda$ = 0.5, and 1.5 ns simulations of both the protein−ligand and free-ligand system for each $\lambda$ value.

Using this approach, we obtained the free-energy estimates listed in Table S2. A histogram of the deviation from the experimental affinities is shown in Figure 2. It can be seen that 54% of the estimates have errors of 4 kJ/mol or less compared to experimental results. The average unsigned error (AUE) over all proteins is 6.0 kJ/mol, whereas the median unsigned error (MUE) is 3.7 kJ/mol. The maximum error is 21 kJ/mol. There is no correlation between the error and how many heavy atoms or how many polar atoms are involved in the transformation. Instead, the performance seems to be more related to what protein is studied. The AUE, MUE, $r^2$, and Kendall's $\tau$ and $\tau_{90}$ are listed in Table 1 for each of the proteins individually.

For CDK2, the AUE is 6 kJ/mol, indicating considerable deviation from the experimental affinities (nine of the 17 transformation had an absolute error larger than 4 kJ/mol), and $r^2$ (0.05) indicates that the predictions are almost random, which is also confirmed by Kendall's $\tau$ (0.2).

The COX2 estimates have an AUE of 10 kJ/mol. The reason for this is that all calculated relative affinities have a smaller magnitude than the experimental ones, except for the two perturbations for which the experimental estimate is less than 1 kJ/mol: The experimental values are −23 to 23 kJ/mol, whereas the calculated ones are −4 to 8 kJ/mol. On the other hand, there is a clear correlation between the calculated and experimental estimates ($r^2$ = 0.4) and eight of the 11 estimates have the correct sign ($\tau$ = 0.5; nine out of 11 of the signs are statistically significant, i.e. $\tau_{90}$ = 0.8). The studied perturbations form two closed cycles, 1 → 2 → 4 → 5 → 7 → 1 and 7 → 8 → 9 → 10 → 7, for which the calculated affinities should vanish exactly, providing a test of the internal consistency of the results. For the first cycle, we obtained a proper result, −1 ± 1 kJ/mol. However, for the second cycle, the results are worse, −3.5 ± 0.3 kJ/mol, indicating that some of the perturbations are not fully converged.


dx.doi.org/10.1021/ci5004027 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

**Table 1. Quality Metrics for the Calculated Affinities for Each of the 10 Studied Proteins Using the Spherical Setup and Three $\lambda$ Values[a]**

| | AUE | MUE | $r^2$ | $\tau$ | $\tau_{90}$ | $nt_{90}$ | $nt_{90}(exp)$ | $nt_{tot}$ |
|---|---|---|---|---|---|---|---|---|
| CDK2 | 6.5 ± 0.4 | 5.6 ± 0.8 | −0.05 ± 0.03 | 0.18 ± 0.15 | 0.43 ± 0.06 | 7 | 10 | 17 |
| COX2 | 10.5 ± 0.5 | 10.7 ± 1.2 | 0.40 ± 0.05 | 0.46 ± 0.13 | 0.78 ± 0.04 | 9 | 9 | 11 |
| DHFR | 2.1 ± 0.4 | 1.7 ± 0.3 | 0.81 ± 0.22 | 0.40 ± 0.17 | 0.60 ± 0.08 | 5 | 6 | 10 |
| ER | 6.4 ± 0.7 | 5.7 ± 1.2 | 0.05 ± 0.06 | −0.33 ± 0.20 | 0.00 ± 0.06 | 4 | 5 | 6 |
| fXa | 4.5 ± 0.5 | 1.5 ± 0.8 | −0.56 ± 0.13 | 0.25 ± 0.22 | 0.00 ± 0.15 | 2 | 3 | 8 |
| Ferritin | 1.6 ± 0.2 | 1.6 ± 0.4 | 0.86 ± 0.05 | 0.71 ± 0.15 | 0.60 ± 0.07 | 5 | 7 | 7 |
| GP | 5.6 ± 0.6 | 4.1 ± 0.9 | 0.09 ± 0.07 | 0.60 ± 0.05 | 0.60 ± 0.05 | 5 | 5 | 5 |
| HIV-PT | 11.0 ± 0.2 | 10.0 ± 0.5 | 0.06 ± 0.03 | −0.25 ± 0.04 | −0.71 ± 0.05 | 7 | 8 | 8 |
| NA | 13.3 ± 1.0 | 13.8 ± 1.4 | 0.81 ± 0.15 | 0.20 ± 0.22 | 0.50 ± 0.03 | 4 | 5 | 5 |
| p38 | 2.1 ± 0.4 | 2.0 ± 0.6 | 0.37 ± 0.13 | 0.14 ± 0.19 | 0.50 ± 0.10 | 4 | 5 | 14 |

[a]The quality metrics are average unsigned error (AUE), the median unsigned error (MUE), the correlation coefficient ($r^2$; a negative sign indicates that $r$ is negative), Kendall's rank correlation coefficient ($\tau$), and $\tau$ calculated only for the transformations for which both the predicted and experimental differences are statistical significantly different from zero at the 90% level. The number of such transformations is indicated by $nt_{90}$, whereas $nt_{90}(exp)$ indicates the number of transformations that have an experimental difference that is statistical significant, and $nt_{tot}$ is the total number of transformations.

The DHFR results are considerably better than for CDK2 and COX2, with an AUE of 2 kJ/mol, a maximum error of 4 kJ/mol, a $r^2$ of 0.8, and a $\tau_{90}$ of 0.6. For this test case, our alchemical perturbation method works well, even if 70% of the calculated estimates are still smaller in magnitude than the experimental ones.

For ER, the results are poor with an AUE of 6 kJ/mol, no correlation, and a negative $\tau = -0.3$ ($\tau_{90} = 0$, reflecting that two of the statistically significant results are correct and two are wrong). Four of the six transformations give absolute errors larger than 4 kJ/mol.

The fXa estimates have an AUE of 5 kJ/mol, but this rather high value comes almost entirely from the 39 → 63 perturbation with an error of 20 kJ/mol; the other seven perturbations give an AUE of only 2 kJ/mol. However, these seven perturbations have quite small experimental binding-affinity differences (less than 5 kJ/mol and only two of them are significantly different from zero at the 90% level), so the correlation is still low ($r^2 = 0.1$).

The ferritin estimates have an AUE of only 2 kJ/mol and an excellent correlation of $r^2 = 0.9$. The maximum error is 3 kJ/mol. Moreover, $\tau = 0.7$ reflects that all except one of the estimates have the correct sign.

The estimates for GP have an AUE of 6 kJ/mol. Again, this is caused by one perturbation, 4 → 1, which gives an error 13 kJ/mol. The other four estimates give an AUE of 4 kJ/mol and a correlation of $r^2 = 0.8$. Only one of the five estimates has an incorrect sign. Also for this target, there is a closed cycle, 1 → 3 → 4 → 1, for which we obtain poor results 6 ± 1 kJ/mol, indicating problems with the convergence (e.g., too few $\lambda$ values).

For HIV-PT, we obtain very poor results: The AUE is 11 kJ/mol, and all perturbations give large errors, 4−21 kJ/mol, even if four of them are simple H → $CH_3$ perturbations. In fact, all calculated estimates are too negative.

The results for NA are similar. All perturbations give a large error of 10−17 kJ/mol, yielding an AUE of 13 kJ/mol, and all the calculated results are too negative. However, since the error is rather systematic, the correlation between the experimental and calculated results is still good, $r^2 = 0.8$.

Finally, p38 gives an AUE of only 2 kJ/mol and a maximum error of 5 kJ/mol. However, most of the experimental differences are small, up to 6 kJ/mol, and only five of them are significantly different from zero at the 90% level. The calculated estimates reproduce the sign of four of these (although only three of the estimates are statistically significant). The correlation is also rather weak, $r^2 = 0.4$.

To summarize, three of the proteins (DHFR, ferritin, and p38) give an AUE better than 4 kJ/mol. Two additional proteins (fXa and GP) give MUE ≤ 4 kJ/mol, indicating that the higher AUE is caused by a single bad result. The other five proteins give quite poor results with AUEs and MUEs of 6−14 kJ/mol. Moreover, only three proteins give correlations above 0.8 (DHFR, ferritin, and NA), because of problems with outliers or that the experiment affinity differences are small. On the other hand, only one protein gives a negative $\tau_{90}$ and all except three give a $\tau_{90}$ that is significantly better than random (which would give $\tau_{90} = 0$). Finally, we note that for 75% of the perturbations, the calculated free-energy difference is smaller than the experimental one.

The statistical uncertainty is also important to consider, and the standard error of all estimates is shown in Table S2. It ranges from 0.05 to 6 kJ/mol with an average of 0.5 kJ/mol, showing that the precision in general is excellent. Only nine of the estimates have a standard error larger than 0.7 kJ/mol, and most of those cases also have a large error in the calculated relative affinities. Clearly, a large standard error indicates incomplete sampling.

**Improving Poor Estimates.** We have seen that the calculated affinities are unsatisfactory for several of the studied proteins. A possible explanation of these poor results is the approximations employed in our alchemical perturbation approach, in particular the use of only three $\lambda$ values and the truncation of the proteins. In a first attempt to improve the results, we tested to insert eight additional $\lambda$ values at 0.1, 0.2, 0.3, 0.4, 0.6, 0.7, 0.8, and 0.9 for CDK2, COX2, ER, GP, HIV-PT, and NA, still using the spherical setup. The individual results are collected in Table S3 and the performance for each protein target is presented in Table 2.

For CDK2, five of the 17 transformations gave a significantly different result with more intermediate states at the 90% level. However, this is mainly due to the high precision of the estimates, and only two transformations gave a difference larger than 2 kJ/mol. For the 30 → 32 transformation, the estimate became 2 kJ/mol more positive, thereby increasing the error compared to experiments to 3 kJ/mol. For the 33 → 21


dx.doi.org/10.1021/ci5004027 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 2. Quality Metrics for the Calculated Affinities for Six Proteins Using the Spherical Setup and 11 $\lambda$ Values (Quality Measures Are the Same As in Table 1)

| | AUE | MUE | $r^2$ | $\tau$ | $\tau_{90}$ | $nt_{90}$ |
|---|---|---|---|---|---|---|
| CDK2 | 6.5 ± 0.3 | 5.7 ± 0.6 | 0.00 ± 0.01 | −0.18 ± 0.15 | 0.43 ± 0.06 | 7 |
| COX2 | 10.4 ± 0.5 | 10.0 ± 1.3 | 0.66 ± 0.06 | 0.64 ± 0.13 | 0.78 ± 0.04 | 9 |
| ER | 7.9 ± 0.7 | 7.6 ± 1.1 | 0.01 ± 0.03 | 0.00 ± 0.14 | 0.60 ± 0.08 | 5 |
| GP | 3.9 ± 0.5 | 3.8 ± 0.8 | 0.15 ± 0.10 | 0.60 ± 0.05 | 0.60 ± 0.05 | 4 |
| HIV-PT | 11.4 ± 0.2 | 11.3 ± 0.6 | 0.27 ± 0.05 | −0.50 ± 0.13 | −0.71 ± 0.05 | 7 |
| NA | 11.3 ± 0.8 | 9.9 ± 1.3 | 0.64 ± 0.14 | 0.60 ± 0.20 | 0.50 ± 0.03 | 4 |

Table 3. Quality Metrics for the Calculated Affinities for Five Proteins Using the Periodic Setup and 13 $\lambda$ Values (Quality Measures Are the Same as in Table 1)

| | AUE | MUE | $r^2$ | $\tau$ | $\tau_{90}$ | $nt_{90}$ |
|---|---|---|---|---|---|---|
| CDK2 | 6.6 ± 0.4 | 5.2 ± 0.7 | 0.27 ± 0.07 | 0.18 ± 0.13 | 0.56 ± 0.05 | 9 |
| COX2 | 7.7 ± 0.5 | 4.6 ± 1.4 | 0.58 ± 0.04 | 0.64 ± 0.13 | 0.78 ± 0.03 | 9 |
| ER | 5.5 ± 0.7 | 3.7 ± 1.0 | 0.78 ± 0.08 | 0.67 ± 0.16 | 1.00 ± 0.09 | 4 |
| HIV-PT | 10.5 ± 0.6 | 10.1 ± 1.4 | 0.00 ± 0.01 | −0.50 ± 0.18 | −0.33 ± 0.05 | 6 |
| NA | 4.7 ± 0.7 | 6.9 ± 1.3 | 0.77 ± 0.12 | 0.60 ± 0.26 | 1.00 ± 0.00 | 3 |

transformation, the estimate changed by 17 kJ/mol, strongly improving the estimate, although it is still 5 kJ/mol from the experimental result. However, besides $\tau_{90}$ (which increases to 0.4), the general performance is not significantly improved.

For COX2, eight of the 11 transformations gave a significantly different result with more intermediate states. However, only two transformations gave a change larger than 2 kJ/mol. These transformations are 1 → 7 and 2 → 3, and one of them was improved and the other became worse, but both still had errors of more than 10 kJ/mol. Yet, $r^2$, $\tau$, and $\tau_{90}$ are all significantly improved (to 0.7, 0.6, and 0.8). The free energies of the two cycles are now both small, −1 kJ/mol.

For ER, all but two of the transformations resulted in significantly different estimates with more intermediate states. Although the differences were larger than 2 kJ/mol for three of these transformations, none of them came closer to experiments. Therefore, only $\tau_{90}$ was significantly improved (to 0.6).

For HIV-PT, seven of the nine transformations resulted in statistically significant differences. Four of them were larger than 2 kJ/mol, but only one of them improved the results compared to experiments. As a consequence, all quality metrics except $r^2$ became worse.

For NA, two of the transformations resulted in significantly different estimates with more intermediate states. The results for transformations 12 → 11 and 17 → 16 became 4 and 5 kJ/mol more positive, i.e. closer to experiment. Therefore, all quality metrics except $r^2$ improved, but the errors were still substantial (6−10 kJ/mol).

For GP, all but one of the transformations resulted in a significant different estimate. For three of these transformations, 3 → 1, 4 → 1, and 4 → 3, the differences were larger than 2 kJ/mol and closer to the experimental value. Hence, the correlation coefficient increased to 0.15, whereas both AUE and MUE decreased to 4 kJ/mol. Thus, GP was the only protein for which the results improved to a satisfactory level with 11 $\lambda$ values. However, one of the perturbations (4 → 1) still gave a large error (10 kJ/mol), and the free energy of the closed cycle (9 ± 1 kJ/mol) indicates that the perturbations are still not converged. A plot of the derivative of the potential energy with respect to $\lambda$ is shown for a typical GP perturbation in Figure S3. It can be seen that there is a very large change between $\lambda$ = 0.0 and 0.1, clearly showing that this system requires more intermediate states at low $\lambda$ values. The situation is remarkably different from the other systems, illustrated by HIV-PT that is plotted in the same figure.

We also tried to run twice as long simulations for COX2 and ER, but the results did not improve. For COX2, only two transformations resulted in a difference larger than 2 kJ/mol, but both transformations still had errors larger than 10 kJ/mol compared to experiments. For ER, none of the transformations gave a statistically significant difference when the simulations were prolonged.

**Results with the Periodic Setup.** It is also possible that the truncation of the protein may deteriorate the calculated affinities. Therefore, we tested a second setup for the five proteins with poor results (CDK2, COX2, ER, HIV-PT, and NA) with simulations of the full protein in a periodic octahedral box, treating long-range electrostatics with Ewald summation using the Amber software and employing 13 $\lambda$ values (the periodic setup).

The results of these calculations are listed in Table S3, and they are summarized in Table 3. It can be seen that the results are in general improved: For the 47 new perturbations, the AUE is reduced from 9 to 7 kJ/mol, the correlation ($r^2$) is improved from 0.03 to 0.29, $\tau$ is improved from 0.11 to 0.28, and $\tau_{90}$ is improved from 0.32 to 0.53. However, the absolute errors decreased for only 62% of the individual perturbations and the maximum error actually increased to 23 kJ/mol.

For ER, the results are now fairly good, with a MUE of 4 kJ/mol, a correlation of 0.8, and a correct sign for the five perturbations that give a significant difference. However, one of the perturbations (4 → 6; involving a OH → H perturbation) still gives a large error (12 kJ/mol).

For NA, the results of all five perturbations are strongly improved, so that AUE decreases from 13 to 5 kJ/mol and $\tau_{90}$ is perfect. However, three of the perturbations still give errors of 7 kJ/mol. The change is essentially a translation of the data, so $r^2$ remains at 0.8.

For COX2, the results for nine of the perturbations are improved and all quality measures become better ($r^2$ to 0.6 and $\tau_{90}$ to 0.8), but six of the transformations still give large errors (5−23 kJ/mol), so the AUE is 8 kJ/mol. The two closed cycles still give free energies of ±1 kJ/mol.


dx.doi.org/10.1021/ci5004027 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

For CDK2, the results of half of the transformations are improved, giving $r^2$ = 0.3 and $\tau_{90}$ = 0.6, but AUE remains at 7 kJ/mol. Finally, HIV-PT still gives poor results with little improvement in any quality metrics (e.g., AUE = 11 kJ/mol and $r^2$ = 0.0).

Interestingly, most perturbations show large differences compared to the previous calculations with an AUD of 7 kJ/mol. A total of 87% of the differences were statistically significant at the 95% level. This is unexpected because previous comparisons between the two approaches for fXa indicated that the two protocols give the same results within 1 kJ/mol. The periodic calculations gave a slightly worse but less varying precision, 0.2−0.9 kJ/mol with an average of 0.6 kJ/mol, except for three HIV-PT transformations (2−3 kJ/mol).

We can directly estimate the effect of the reduced number of $\lambda$ values by calculating the relative binding affinities from the new calculations based on only three $\lambda$ values (0.0, 0.5, and 1.0). As can be seen in Tables S3 and S4, the results deteriorate: For example, AUE increases to 7−23 kJ/mol for the five proteins (11 kJ/mol for all). The individual estimates change by up to 51 kJ/mol with an AUD of 7 kJ/mol. For 51% of the estimates, the change is statistically significant at the 95% level (i.e., typically 3 kJ/mol) and for 83% of the estimates, the results deteriorate. The two closed cycles for COX2 give very poor results (6 and 36 kJ/mol), showing severe problems with the convergence, much worse than for the spherical setup. Clearly, using only three $\lambda$ values is a poor approximation, especially for the periodic simulations, and cannot be recommended. Finally, it should be admitted that it is somewhat risky to rerun only proteins that gave poor results with improved methods—it is possible that the other systems gave good results only by chance, although previous calculations with fXa have shown good results both with accurate and more approximate approaches.[9,10]

**What Performance Can Be Expected?** When discussing the accuracy of estimated binding affinities, it is important to remember that both the experimental and calculated affinities have a limited precision. Therefore, even if the calculations gave the correct results (within this precision), $r^2$ and $\tau$ will not be 1.0.[46,47] This can be illustrated by a statistical simulation in which we assume that the predicted affinities give the same results as the experimental affinities, but both are affected by a normal-distributed statistical uncertainty. The experimental uncertainties were 0.1−2.2 kJ/mol, but they were not reported for five of the proteins. Therefore, we assume an experimental uncertainty of 1.5 kJ/mol for all data.[46] For the calculated affinities, we assume a standard deviation of 0.5 kJ/mol, which is roughly the average uncertainty of the BAR predictions in this study.

With these uncertainties, we sampled 10 000 series of experimental and such exact "estimated" affinities and calculated the $r^2$ and $\tau$ between the two sets. The averages of these 10 000 series are presented in Table 4 and represent the best possible results given the uncertainties in the experiments and predictions. Comparing these results with the BAR results in Tables 1−3, it can be seen that the BAR results for ferritin are actually better than what can be expected. This is partly because the experimental uncertainty for ferritin is only 0.18 kJ/mol. Using this uncertainty in the estimation of the optimum performance, the observed $r^2$ is more realistic, although the observed $\tau$ is still larger than the calculated optimum (which of course is possible, because the BAR result is only one possible outcome of a random sampling, giving rise to the average results in Table 4). It can also be seen that $r^2$ for the BAR results of DHFR and $\tau$ for ER approach the optimum values, showing that those predictions are close to ideal, although the quality metrics themselves are not perfect. The $r^2$ for ER and NA, as well as $\tau$ for GP and NA, are also within 0.14 of the ideal values, whereas for the other systems, the results are clearly far from ideal.

**Table 4. Estimated Optimum $r^2$ and $\tau$[a]**

| | optimum | | BAR | |
|---|---|---|---|---|
| | $r^2$ | $\tau$ | $r^2$ | $\tau$ |
| CDK2 | 0.90 | 0.78 | 0.27 | 0.18 |
| COX2 | 0.99 | 0.88 | 0.58 | 0.64 |
| DHFR | 0.86 | 0.70 | 0.81 | 0.40 |
| ER | 0.91 | 0.66 | 0.78 | 0.67 |
| fXa | 0.86 | 0.71 | −0.56 | 0.25 |
| Ferritin | 0.74 | 0.51 | 0.86 | 0.71 |
| GP | 0.84 | 0.71 | 0.15 | 0.60 |
| HIV-PT | 0.86 | 0.69 | 0.00 | −0.50 |
| NA | 0.91 | 0.73 | 0.77 | 0.60 |
| p38 | 0.76 | 0.66 | 0.37 | 0.14 |

[a]The average metric was estimated from 10 000 resamples of the experimental free energies, assuming a normal distribution with uncertainties of 0.5 and 1.5 kJ/mol for the predicted and experimental free energies, respectively. Our best estimated BAR results are also included (CDK2, COX2, ER, HIV-PT, and NA from Table 3, for GP from Table 2, and for the other systems from Table 1).

Some of the proteins in this study have been used previously in theoretical studies. CDK2, COX2, and NA were studied by Essex and co-workers to test if implicit-solvation simulations could be used as a tool in lead optimization.[11] For CDK2, they obtained poor predictions with $r^2$ = 0.09−0.16 and AUE = 13−19 kJ/mol. For COX2, they got decent correlations with $r^2$ = 0.70−0.85 and AUE = 3−5 kJ/mol. Finally, for NA, they obtained $r^2$ = 0.8 and MUE = 5−14 kJ/mol. However, it should be noted that they used a larger set of transformations than in this study, so the results are not directly comparable. The p38 system has been used in several studies.[12,13] Pearlman and Charifson reported a predictive index of 0.8 using TI, but only if the protein was restrained to the crystal structure, indicating poor sampling.[12] Jorgensen and co-workers reported a predictive index of 0.4 and AUE of 7 kJ/mol. However, they were able to improve this by including explicit water molecules in their binding site.[13] DHFR was used in a study with the MM/PB(GB)SA approach (molecular mechanics with Poisson−Boltzmann or generalized Born and surface-area solvation), and they obtained excellent correlation with experimental data with $r^2$ > 0.8 for most of their tested protocols.[14] The HIV-PT system has been used in some studies with heavily fitted linear interaction energy models.[15,16] Ferritin and fXa have been used in several studies by us, especially when testing the MM/GBSA approach.[9,10,17] For CDK2 and fXa, relative binding affinities have been calculated by TI and BAR for other large sets of ligand.[48] For fXa, a decent $r^2$ = 0.58 but a rather high root-mean-squared error (RMSE) of 9 kJ/mol were obtained. For CDK2, they obtained a similar RMSE, but the $r^2$ was only 0.12. This could be improved to 0.36 if eight transformations with large changes in the protein structure were omitted. They also obtained results with similar or better quality with the MM/PBSA approach (but not with the linear interaction energy approach), but only for selected subsets of ligands or variations in the method (structures from MD or


dx.doi.org/10.1021/ci5004027 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 5. Diagnostic Measures of the Performance of the Various Transformations: The Bhattacharyya Coefficient for the Energy Distribution Overlap ($\Omega$),[49] the Wu and Kofke Overlap Measures of the Energy Probability Distributions ($K_{AB}$) and Their Bias Metrics ($\Pi$),[50,51] the Weight of the Maximum Term in the Exponential Average ($w_{max}$), the Difference of the Forward and Backward Exponential Average Estimate ($\Delta\Delta G_{EA}$), and the Difference between the BAR and TI Estimates ($\Delta\Delta G_{TI}$)[a]

| | $\Omega$ | $K_{AB}$ | $\Pi$ | $w_{max}$ | $\Delta\Delta G_{EA}$ | $\Delta\Delta G_{TI}$ | $\Delta\Delta G_{exp}$ |
|---|---|---|---|---|---|---|---|
| CDK2 | | | | | | | |
| 22→21 | 0.73 | 0.82 | 0.80 | 0.10 | **4.3** | 0.4 | 2.7 |
| 23→21 | 0.78 | 0.86 | 1.10 | 0.06 | 1.8 | 0.1 | 5.7 |
| 24→21 | 0.76 | 0.89 | 1.00 | 0.06 | 0.7 | 0.0 | 16.7 |
| 25→21 | 0.78 | 0.86 | 1.10 | 0.08 | 0.8 | 0.2 | 0.5 |
| 26→21 | 0.76 | 0.89 | 1.20 | 0.04 | 0.6 | 0.1 | 1.1 |
| 27→21 | 0.75 | 0.92 | 0.80 | 0.10 | 0.8 | 0.0 | 2.3 |
| 28→21 | 0.77 | 0.83 | 0.90 | 0.09 | 0.6 | 0.1 | 13.6 |
| 29→21 | 0.76 | 0.89 | 1.10 | 0.09 | 1.6 | 0.8 | 4.7 |
| 30→21 | 0.72 | 0.85 | 1.00 | 0.07 | 2.5 | 0.1 | 10.8 |
| 30→32 | 0.74 | 0.85 | 0.70 | **0.22** | 1.4 | 0.2 | 1.0 |
| 31→32 | 0.74 | 0.87 | 0.90 | 0.08 | 1.7 | 0.1 | 5.1 |
| 33→21 | 0.80 | 0.88 | 0.80 | 0.13 | 1.3 | 0.3 | 9.7 |
| 33→35 | 0.76 | 0.87 | **0.40** | **0.44** | **4.2** | 0.8 | 7.6 |
| 34→35 | 0.74 | 0.83 | 0.90 | 0.10 | 3.3 | 0.0 | 3.5 |
| 36→21 | 0.78 | 0.88 | 1.00 | 0.10 | 1.3 | 0.0 | 20.1 |
| 37→21 | 0.79 | 0.87 | 1.10 | 0.06 | 1.0 | 0.0 | 0.3 |
| 38→32 | **0.57** | **0.23** | 0.50 | 0.15 | 9.3 | 1.3 | 6.2 |
| COX2 | | | | | | | |
| 1→7 | 0.77 | 0.87 | 0.50 | 0.07 | 2.6 | **5.1** | 10.3 |
| 2→1 | **0.37** | **0.11** | 1.00 | 0.04 | **6.7** | 0.4 | 4.6 |
| 2→3 | **0.43** | **0.15** | **0.00** | 0.10 | 3.4 | **7.8** | 1.4 |
| 2→4 | 0.90 | 0.89 | 1.10 | 0.04 | 0.9 | 0.1 | 8.6 |
| 5→4 | 0.79 | 0.84 | 1.30 | 0.04 | 0.5 | 0.2 | 0.4 |
| 5→7 | 0.77 | 0.90 | 0.80 | 0.18 | 1.2 | 0.7 | 13.2 |
| 6→1 | 0.79 | 0.82 | 1.20 | 0.05 | 1.8 | 0.5 | 2.1 |
| 7→8 | 0.77 | 0.89 | 1.10 | 0.05 | 1.3 | 0.6 | 19.0 |
| 7→10 | 0.75 | 0.85 | 1.10 | 0.03 | 0.9 | 0.7 | 22.9 |
| 9→8 | 0.74 | 0.76 | 1.10 | 0.07 | 0.7 | 0.1 | 0.9 |
| 10→9 | 0.76 | 0.83 | 1.00 | 0.05 | 1.1 | 0.2 | 1.5 |
| ER | | | | | | | |
| 1→3 | **0.57** | **0.32** | **0.10** | **0.38** | **5.5** | 0.5 | 3.2 |
| 2→1 | 0.76 | 0.87 | 1.00 | 0.08 | 1.4 | 0.8 | 3.6 |
| 3→6 | 0.74 | 0.87 | 0.80 | 0.09 | 2.3 | 0.4 | 7.2 |
| 4→5 | 0.74 | 0.86 | 1.10 | 0.06 | 1.1 | 0.5 | 3.8 |
| 4→6 | 0.75 | 0.77 | 1.00 | 0.07 | 1.8 | 0.4 | 11.7 |
| 8→4 | 0.76 | 0.90 | 0.60 | **0.28** | 2.8 | 1.6 | 3.7 |
| HIV-PT | | | | | | | |
| 2→1 | 0.72 | 0.87 | 0.70 | 0.08 | 3.0 | 0.1 | 13.2 |
| 3→1 | 0.74 | 0.83 | 0.60 | **0.31** | **4.1** | 0.4 | 16.9 |
| 7→6 | 0.78 | 0.84 | **-0.10** | **0.38** | **103.9** | 1.2 | 10.1 |
| 8→6 | 0.73 | 0.86 | **-0.10** | **0.29** | **104.0** | 0.4 | 13.3 |
| 12→11 | 0.76 | 0.88 | **-0.10** | **0.22** | **116.9** | **5.1** | 1.6 |
| 13→11 | 0.73 | 0.94 | 0.60 | **0.29** | **5.6** | 0.3 | 16.4 |
| 22→21 | 0.78 | 0.90 | 1.10 | 0.09 | 0.8 | 0.0 | 3.3 |
| 23→21 | 0.79 | 0.94 | 0.60 | 0.18 | **5.5** | 0.0 | 9.4 |
| NA | | | | | | | |
| 12→11 | 0.78 | 0.84 | 0.90 | 0.11 | 1.9 | 0.0 | 7.2 |
| 13→11 | 0.76 | 0.87 | 0.80 | 0.08 | **4.8** | 0.4 | 1.8 |
| 14→12 | 0.81 | 0.86 | 0.80 | 0.13 | 1.5 | 0.0 | 7.0 |
| 16→15 | 0.79 | 0.89 | 0.60 | 0.09 | 2.5 | 0.2 | 0.8 |
| 17→16 | 0.77 | 0.83 | 0.70 | 0.20 | 2.2 | 0.1 | 6.9 |
| $r$ | 0.13 | 0.19 | −0.03 | 0.12 | 0.04 | −0.12 | |

[a]The table shows the maximum ($w_{max}$, $\Delta\Delta G_{EA}$, $\Delta\Delta G_{TI}$) or minimum ($\Omega$, $K_{AB}$, $\Pi$) values of these estimates over the 12 + 12 individual perturbations and possibly for both forward and backward perturbations ($\Omega$, $K_{AB}$, $\Pi$, $w_{max}$), for the complex and free-ligand transformations for the periodic simulations with 13 $\lambda$ values. Problematic perturbations are marked in bold face, i.e. those with $\Omega$ < 0.7, $K_{AB}$ < 0.7, $\Pi$ < 0.5,[50] $w_{max}$ > 0.2, $\Delta\Delta G_{EA}$ > 4 kJ/mol, or $\Delta\Delta G_{TI}$ > 4 kJ/mol. In addition, the absolute difference between the BAR and experimental estimate of the binding affinity ($\Delta\Delta G_{exp}$ in kJ/mol) is included, as well as the correlation between the various measures and $\Delta\Delta G_{exp}$ ($r$; on the last line).

minimization; variations in the dielectric constant). A recent large-scale test of alchemical perturbation methods on five other proteins gave similar results, with RMSE = 1−9 kJ/mol and $r^2$ = 0.25−0.96, depending on the protein target.[40]

**Analysis of the Performance.** In this section, we try to explain why the alchemical perturbations were successful on some systems but failed on others. One possibility is that this is connected to the performance of the alchemical perturbations. We tested several diagnostic tools suggested in the literature, measuring the overlap of the distributions or the convergence of the perturbations, viz., the Bhattacharyya coefficient for the energy distribution overlap ($\Omega$),[49] the Wu and Kofke overlap measures of the energy probability distributions ($K_A$) and their bias metrics ($\Pi$),[50,51] the weight of the maximum term in the exponential average ($w_{max}$), the difference of the forward and backward exponential average estimate ($\Delta\Delta G_{EA}$), and the difference between the BAR and TI estimates ($\Delta\Delta G_{TI}$, although this difference may also reflect the integration error in TI[52]). The maximum ($w_{max}$, $\Delta\Delta G_{EA}$, $\Delta\Delta G_{TI}$) or minimum ($\Omega$, $K_{AB}$, $\Pi$) values of these estimates over the 12 + 12 individual perturbations for the complex and free-ligand transformations are listed in Table 5 for the periodic simulations with 13 $\lambda$ values.

After some testing, we decided to use the following criteria for problematic perturbations: $\Omega$ < 0.7, $K_{AB}$ < 0.7, $\Pi$ < 0.5,[50] $w_{max}$ > 0.2, $\Delta\Delta G_{EA}$ > 4 kJ/mol, or $\Delta\Delta G_{TI}$ > 4 kJ/mol. It can be seen that there is some consistency between the various measures, pointing out the same transformations as problematic. However, $w_{max}$ and especially $\Delta\Delta G_{EA}$ often point out too many transformations (the exponential averaging is often problematic in the direction from the smaller to the larger ligand, but not in the opposite direction, something BAR should be able to cure), whereas $\Omega$, $K_{AB}$, and $\Delta\Delta G_{TI}$ point out too few. In total, 10 transformations were pointed out by at least two diagnostics to be problematic, mainly for HIV-PT, indicating that more $\lambda$ values or longer simulations are needed.

Unfortunately, there is very little correlation between the diagnostics and the error of the BAR results compared to experimental data ($r$ < 0.2), and none of the criteria could correctly predict whether the transformation is successful (error <4 kJ/mol) in more than half of the transformations (19−24 correct predictions, out of 47, i.e. similar to a random guess). However, when applied to the calculations with three $\lambda$ values,


dx.doi.org/10.1021/ci5004027 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 6. Average Properties of the Ligands and the Protein−ligand Complexes As Well As Their Correlation to the Average Performance of the BAR Calculations (MUE and $\tau_{90}$; for CDK2, COX2, ER, HIV-PT, and NA, Metrics Taken from Table 3, for GP from Table 2, and for the Other Systems from Table 1; Values in Brackets Show the Correlations Omitting HIV-PT)

| | #H-bond donors | #H-bond acceptors | molecular weight | #rotatable bonds | ΔSASA[a] | resolution | RMSF[b] |
|---|---|---|---|---|---|---|---|
| CDK2 | 1.6 | 5.3 | 313.5 | 3.4 | 64.5 | 2.3 | 0.28 |
| COX2 | 1.2 | 5.3 | 395.8 | 4.5 | 74.3 | 3.0 | 0.27 |
| DHFR | 2.9 | 4.6 | 237.6 | 1.6 | 96.4 | 2.3 | 0.32 |
| ER | 2.7 | 5.1 | 461.4 | 6.0 | 63.1 | 1.9 | 0.35 |
| fXa | 5.1 | 7.5 | 446.6 | 8.1 | 97.2 | 2.0 | 0.33 |
| Ferritin | 1.0 | 1.0 | 154.9 | 1.8 | 68.8 | 1.9 | 0.29 |
| GP | 5.8 | 9.8 | 268.6 | 1.4 | 67.7 | 2.4 | 0.23 |
| HIV-PT | 5.5 | 10.5 | 551.6 | 19.4 | 50.0 | 2.5 | 0.34 |
| NA | 4.3 | 8.9 | 296.3 | 6.6 | 60.0 | 2.0 | 0.26 |
| p38 | 0.1 | 4.1 | 432.9 | 3.0 | 53.4 | 2.5 | 0.30 |
| $r$(MUE) | 0.42 | 0.67 | 0.45 | 0.76 | −0.61 | 0.26 | −0.01 |
| | (0.16) | (0.51) | (0.03) | (0.27) | (−0.49) | (0.14) | (−0.43) |
| $r(\tau_{90})$ | −0.41 | −0.37 | −0.48 | −0.68 | −0.04 | −0.15 | −0.55 |
| | (−0.16) | (0.01) | (−0.12) | (−0.07) | (−0.57) | (0.02) | (−0.21) |

[a]ΔSASA = (SASA(bound) − SASA(free))/SASA(free) in percentage, where SASA is the solvent-accessible surface area of the ligand when it is free in solution or when it is bound to the receptor. It was estimated from the starting structure of the simulations. [b]The average root-mean-square fluctuation (RMSF) of the backbone CA atoms (in Å), estimated from the simulations at $\lambda$ = 0.0.

significantly more transformations are predicted to be problematic, showing that these criteria can be used to decide if too few $\lambda$ values are used.

Next, we have analyzed the structural and chemical features of the proteins and ligands. Such an analysis has previously been performed in the context of virtual screening.[53] For the ligands, we calculated the number of hydrogen-bond donors and acceptors, molecular weight, and number of rotatable bonds (see Table 6). We then computed correlation between the average properties for each protein and the BAR performance (MUE and $\tau_{90}$). The number of hydrogen-bond donors and acceptors as well as the molecular weight showed a rather weak correlation to the BAR performance (0.37−0.48), except the number of hydrogen-bond acceptors, which showed a somewhat stronger correlation to MUE (0.67). The number of rotatable bonds showed an even stronger correlation to both MUE and $\tau_{90}$ (0.76 and −0.68). However, all these correlations come mainly from HIV-PT, which showed the worst performance and also the largest ligands with by far the largest number of rotable bonds, but also many hydrogen-bond donors and acceptors. If HIV-PT is omitted, all correlations drop to 0.27 or less, except for the number of hydrogen-bond acceptors vs MUE (0.51). This shows that the poor results for HIV-PT can at least partly be explained by its very large and flexible ligands, making sampling problematic (although for relative affinities we would expect a fair degree of cancellation of this effect, unless the ligands have different binding modes). However, for the other targets, we see no consistent correlation to the properties of the ligands.

Next, we studied the protein−ligand complexes. We first computed how buried the ligands are in the binding pocket by taking the difference of the solvent-accessible surface of the ligand when free in solution and bound to the receptor (ΔSASA in Table S5). ΔSASA showed some anticorrelation to the MUE (−0.61) but no correlation to the ranking. This reflects that the proteins with the most buried ligands (DHFR and fXa) gave good MUEs, whereas HIV-PT with the least buried ligands gave the worst results. This probably reflects that buried ligands show smaller dynamics than exposed ones, reducing the need of conformational sampling. However, the correlation is far from perfect—p38 has almost as exposed ligands as HIV-PT, but gave a low MUE, whereas COX2 with the third highest ΔSASA gave poor results.

It is reasonable to expect that the quality of the crystal structure may affect the alchemical perturbation results. In this study, we have employed structures with resolutions of 1.9−3.0 Å (Table 6). However, there was no correlation between the resolution of the crystal structure and the accuracy of the BAR results ($r$ < 0.3). Likewise, we have computed the root-mean-squared fluctuation (RMSF) for the protein backbone in the MD simulations (see Table 6). However, the 10 proteins gave quite similar results (0.23−0.35 Å), and there was no consistent correlation to the BAR results ($r$ = −0.01 and −0.43).

Finally, we note that the experimental affinities for ferritin were determined by isothermal calorimetry,[54] whereas they were obtained from various biochemical assays for the other proteins.[12,55−63] For COX2, ER, NA, and p38, only $IC_{50}$ estimates were reported, whereas for the other proteins, $K_i$ values were measured. It is notable that three of the former proteins gave poor BAR results, indicating that the accuracy of experimental data might affect the results.

**Docking Calculations.** To put the alchemical perturbation results in a proper perspective, we also tested to study the same 91 transformations with docking calculations. We tested two different widely used docking software, GOLD 5.1[41] with the ChemScore[42,43] scoring function and Dock 6.5[44] with grid scoring. The results of these calculations are collected in Table S5 and summarized in Table 7. Since the docking calculations do not give energies or any uncertainties of the estimated scores, we can only compare $r^2$ and $\tau$. From Table 7, it can be seen that GOLD gave significantly (by at least 0.1) better $r^2$ than BAR for COX2 (0.6; $r^2$ for fXa is also better, but $r^2$ = 0.1 indicates essentially no correlation). On the other hand, it gave worse results for four proteins, CDK2, DHFR, ferritin, and p38 (as before, we use results for CDK2, COX2, ER, HIV-PT, and NA from Table 3, for GP from Table 2, and for the others from Table 1). Likewise, GOLD gave a better $\tau$ for one target, HIV-PT (0.75), but worse results for six proteins (CDK2, COX2, DHFR, ER, GP, and p38).


dx.doi.org/10.1021/ci5004027 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 7. Quality Measures for Affinities Obtained by Docking for Each of the 10 Studied Proteins[a]

| | GOLD, best | | GOLD, alt. | | Dock, best | | Dock, alt. | |
|---|---|---|---|---|---|---|---|---|
| | $r^2$ | $\tau$ | $r^2$ | $\tau$ | $r^2$ | $\tau$ | $r^2$ | $\tau$ |
| CDK2 | −0.04 | −0.06 | 0.00 | 0.06 | 0.00 | −0.18 | 0.00 | −0.18 |
| COX2 | 0.63 | 0.45 | 0.63 | 0.45 | −0.11 | 0.27 | −0.05 | 0.09 |
| DHFR | 0.69 | 0.20 | 0.69 | 0.20 | 0.55 | 0.60 | 0.04 | −0.20 |
| ER | 0.89 | 1.00 | 0.84 | 1.00 | 0.02 | 0.00 | 0.09 | 0.00 |
| fXa | 0.11 | 0.00 | 0.00 | 0.00 | 0.00 | −0.50 | 0.01 | −0.50 |
| Ferritin | 0.57 | 0.71 | 0.45 | 0.71 | 0.00 | 0.14 | −0.13 | 0.14 |
| GP | 0.07 | −0.60 | 0.07 | −0.60 | −0.15 | −0.20 | −0.12 | 0.20 |
| HIV-PT | 0.06 | 0.75 | 0.00 | 0.25 | 0.00 | 0.50 | 0.00 | 0.25 |
| NA | 0.88 | 0.60 | 0.88 | 0.60 | −0.57 | 0.60 | −0.55 | 0.20 |
| p38 | −0.03 | −0.21 | −0.11 | −0.07 | 0.00 | −0.29 | 0.10 | −0.21 |

[a]Two docking methods were used (GOLD and Dock), and for each, two results are given. The first (best) is the best score among all obtained structures. In the second (alt.), the best score is reported that has an RMSD below 2 Å or the score of the pose with the lowest RMSD, if no poses with RMSD < 2 Å were found. The raw data are collected in Table S5.

Dock gave somewhat worse results: It gave an improved $\tau$ for DHFR and HIV-PT (0.6 and 0.5, respectively) but worse results for all the other targets except NA. The correlation coefficient was worse for all targets except HIV-PT and fXa (and for the latter two targets, Dock gave no correlation, $r^2$ = 0.0).

We also calculated the root-mean-squared deviation (RMSD) of the docked poses from the starting structure (crystal or modeled structure). For many proteins, this deviation is small, meaning that the docked pose is close to the expected binding position. However, for some proteins, large deviations were observed for all (ferritin and HIV-PT) or most (p38) of the poses. We then tried to instead use poses with an RMSD below 2 Å or the pose with the lowest RMSD, if no poses with a RMSD < 2 Å was found. However, this did not change the results significantly.

Therefore, we can conclude that for the tested transformations, the alchemical perturbations clearly give better results than docking with these two software. Of course, this comes with a much higher computational effort for the BAR calculations and the fact that we only estimate relative affinities with BAR, whereas the docking gives absolute affinities also.

## CONCLUSIONS

In this paper, we have studied the performance of alchemical perturbation simulations, using the BAR approach, to estimate relative binding affinities for a large and diverse test set involving 107 ligands binding to 10 different proteins (91 relative affinities in total). In particular, we wanted to test our recently suggested approach to speed up the calculations by using only a single intermediate state and spherical systems where the protein outside 20 Å from a central atom of the ligand is ignored.[9,10]

In general, the results are rather good: 54% of the calculated affinity differences agree with experimental data to within 4 kJ/mol, which is a reasonable target accuracy (the reported standard error of the experimental affinities is 0.1−2.2 kJ/mol). The precision is also excellent, with a median value for the standard error of only 0.3 kJ/mol—only seven of the calculated affinities have a standard error above 1 kJ/mol. Moreover, the ranking of the ligands is better than random for seven of the proteins. However, for 46% of the affinities, the results are less satisfactory. Therefore, we tried to use more intermediate states or include the entire protein in the calculations using periodic boundaries. This resulted in improved results, although, for CDK2, COX2, and HIV-PT, the results are still poor.

The present calculations are fairly fast and automatic. For example, the spherical calculations for fXa with 3 or 11 $\lambda$ values took 36 CPU h on six or 22 processors in total for one transformation.[9] The corresponding periodic simulations took 56 h on six or 26 processors with 3 or 13 $\lambda$ values.[10] Given ligand−protein input structures, the setup and run of the simulations are essentially automatic, using the scripts described in the Methods section.

From this study, we can conclude that alchemical perturbations with BAR is a promising approach for the calculation of relative binding affinities in a lead-optimization setting, giving better results than docking calculations with GOLD and DOCK. However, it performs poorly for some targets, probably owing to conformational changes during the binding or problems with the molecular-mechanics force field for some types of ligands or interactions. In particular, it seems to have a problem with large and flexible ligands and solvent-exposed binding sites, e.g., for HIV-PT. Moreover, even for well-behaving proteins, a few transformations gave large errors, probably indicating a change in the binding mode or the involvement of slowly equilibrating water molecules in the binding. It is a future challenge to solve these problems to make this approach more robust. Problems with conformational changes could in principle be solved by longer simulations[64−66] or by accelerated-sampling methods,[67−74] but this would typically be too expensive in this large-scale test and in a drug-design workflow. Our results clearly show that employing only a single intermediate state in the perturbations to speed up the calculations works only for a few proteins and is not a general approach. Likewise, it seems to be advisable to include the full protein in the simulations.

## ASSOCIATED CONTENT

### Supporting Information

Binding free energies estimated by thermodynamic integration and exponential averaging, details of the setup of the proteins, the studied transformations and the raw binding free energies obtained for each setup, quality metrics for the calculated affinities for the periodic setup and 3 $\lambda$ values, raw results for all the docking calculations, ligand-interaction diagrams and active-site sketches for each protein, as well as the derivative of the potential energy with respect to $\lambda$ for a typical perturbation for


dx.doi.org/10.1021/ci5004027 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

the GP and HIV-PT systems. This material is available free of charge via the Internet at http://pubs.acs.org.

## AUTHOR INFORMATION

### Corresponding Author

*E-mail: Ulf.Ryde@teokem.lu.se. Tel: +46−46 222 45 02. Fax: +46−46 222 86 48.

### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

This investigation has been supported by grants from the Swedish research council (grant 2010-5025), from the FLÄK research school in pharmaceutical science at Lund University, and the Knut and Alice Wallenberg Foundation (KAW 2013.0022). It has also been supported by computer resources of Lunarc at Lund University, NSC at Linköping University, C3SE at Chalmers University of Technology, and HPC2N at Umeå University.

## REFERENCES

(1) Michel, J.; Essex, J. W. Prediction of Protein-Ligand Binding Affinity by Free Energy Simulations: Assumptions, Pitfalls and Expectations. *J. Comput. Aided Mol. Des.* **2010**, *24*, 639−658.

(2) Chipot, C.; Rozanska, X.; Dixit, S. B. Can Free Energy Calculations Be Fast and Accurate at the Same Time? Binding of Low-Affinity, Non-Peptide Inhibitors to the Sh2 Domain of the Src Protein. *J. Comput. Aided Mol. Des.* **2005**, *19*, 765−770.

(3) Jorgensen, W. L. Efficient Drug Lead Discovery and Optimization. *Acc. Chem. Res.* **2009**, *42*, 724−733.

(4) Gohlke, H.; Klebe, G. Approaches to the Description and Prediction of the Binding Affinity of Small Molecule Ligands to Macromolecular Receptors. *Angew. Chem., Int. Ed.* **2002**, *41*, 2644−2676.

(5) Foloppe, N.; Hubbard, R. Towards Predictive Ligand Design with Free-Energy Based Computational Methods? *Curr. Med. Chem.* **2006**, *13*, 3583−3608.

(6) Warshel, A. Energetics of Enzyme Catalysis. *Proc. Natl. Acad. Sci. U. S. A.* **1978**, *75*, 5250−5254.

(7) Warshel, A.; Sussman, F.; King, G. Free Energy of Charges in Solvated Proteins: Microscopic Calculations Using a Reversible Charging Process. *Biochemistry* **1986**, *25*, 8368−8372.

(8) Jones-Hertzog, D. K.; Jorgensen, W. L. Binding Affinities for Sulfonamide Inhibitors with Human Thrombinusing Monte Carlo Simulations with a Linear Response Method. *J. Med. Chem.* **1997**, *40*, 1539−1549.

(9) Genheden, S.; Ryde, U. Improving Efficiency of Protein−Ligand Binding Free-Energy Calculations by System Truncation. *J. Chem. Theory Comput.* **2012**, *8*, 1449−1458.

(10) Genheden, S.; Nilsson, I.; Ryde, U. Binding Affinities of Factor Xa Inhibitors Estimated by Thermodynamic Integration and MM/GBSA. *J. Chem. Inf. Model.* **2011**, *51*, 947−958.

(11) Liu, S.; Wu, Y.; Lin, T.; Abel, R.; Redmann, J. P.; Summa, C. M.; Jaber, V. R.; Lim, N. M.; Mobley, D. L. Lead Optimization Mapper: Automating Free Energy Calculations for Lead Optimization. *J. Comput. Aided Mol. Des.* **2013**, *27*, 755−770.

(12) Pearlman, D. A.; Charifson, P. S. Are Free Energy Calculations Useful in Practice? A Comparison with Rapid Scoring Functions for the p38 MAP Kinase Protein System. *J. Med. Chem.* **2001**, *44*, 3417−3423.

(13) Lucarelli, J.; Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. Effects of Water Placement on Predictions of Binding Affinities for p38 MAPalpha Kinase Inhibitors. *J. Chem. Theory Comput.* **2010**, *6*, 3850−3856.

(14) Rastelli, G.; Del Rio, A.; Degliesposti, G.; Sgobba, M. Fast and Accurate Predictions of Binding Free Energies Using MM-PBSA and MM-GBSA. *J. Comput. Chem.* **2010**, *31*, 797−810.

(15) Huang, D.; Caflisch, A. Efficient Evaluation of Binding Free Energy. *J. Med. Chem.* **2004**, *47*, 5791−5797.

(16) Chen, S.-L.; Zhao, D.-X.; Yang, Z.-Z. An Estimation Method of Binding Free Energy in Terms of ABEEM$\sigma\pi$/MM and Continuum Electrostatics Fused Into Lie Method. *J. Comput. Chem.* **2011**, *32*, 338−348.

(17) Mikulskis, P.; Genheden, S.; Wichmann, K.; Ryde, U. A Semiempirical Approach To Ligand-Binding Affinities: Dependence on The Hamiltonian and Corrections. *J. Comput. Chem.* **2012**, *33*, 1179−1189.

(18) Michel, J.; Verdonk, M. L.; Essex, J. W. Protein-Ligand Binding Affinity Predictions by Implicit Solvent Simulations: A Tool for Lead Optimization? *J. Med. Chem.* **2006**, *49*, 7427−7439.

(19) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B. P.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossvai, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, Q.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. *AMBER 11*; University of California: San Francisco, CA, 2010.

(20) 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. Bioinform.* **2006**, *65*, 712−725.

(21) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.

(22) Wang, J. M.; Wolf, R. M.; Caldwell, K. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.

(23) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera–A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25*, 1605−1612.

(24) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. Application of RESP Charges To Calculate Conformational Energies, Hydrogen Bond Energies, and Free Energies of Solvation. *J. Phys. Chem.* **1993**, *97*, 10269−10280.

(25) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. The Development and Use of Quantum Mechanical Molecular Models. 76. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(26) Besler, B. H.; Merz, K. M.; Kollman, P. A. Atomic Charges Derived from Semiempirical Methods. *J. Comput. Chem.* **1990**, *11*, 431−439.

(27) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impley, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926−935.

(28) Tembe, B. L.; McCammon, J. A. Ligand-Receptor Interactions. *Comp. Chem.* **1984**, *8*, 281−283.

(29) Bennett, C. H. Efficient Estimation of Free Energy Differences from Monte Carlo Data. *J. Comput. Phys.* **1976**, *22*, 245−268.

(30) Beutler, T. C.; Mark, A. E.; Van Schaik, R. C.; Gerber, P. R.; Van Gunsteren, W. F. Avoiding Singularities and Numerical Instabilities in Free Energy Calculations Based on Molecular Simulations. *Chem. Phys. Lett.* **1994**, *222*, 529−539.

(31) Zacharias, M.; Straatsma, T. P.; McCammon, J. A. Separation-Shifted Scaling, a New Scaling Method for Lennard- Jones Interactions in Thermodynamic Integration. *J. Chem. Phys.* **1994**, *100*, 9025−9031.

(32) Marelius, J.; Kolmodin, K.; Feierberg, I.; Åqvist, J. Q: A Molecular Dynamics Program for Free Energy Calculations and Empirical Valence Bond Simulations in Biomolecular Systems. J. *J. Mol. Graph. Model.* **1998**, *16*, 213−225.

(33) Steinbrecher, T.; Mobley, D. L.; Case, D. A. Nonlinear Scaling Schemes for Lennard-Jones Interactions in Free Energy Calculations. *J. Chem. Phys.* **2007**, *127*, 214108.
(34) 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.
(35) Lee, F. S.; Warshel, A. A Local Reaction Field Method for Fast Evaluation of Long-range Electrostatic Interactions in Molecular Simulations. *J. Chem. Phys.* **1992**, *97*, 3100−3107.
(36) 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.
(37) Kaus, J. W.; Pierce, L. T.; Walker, R. C.; McCammon, J. A. Improving the Efficiency of Free Energy Calculations in the Amber Molecular Dynamics Package. *J. Chem. Theory Comput.* **2013**, *9*, 4131−4139.
(38) Wu, X.; Brooks, B. R. Self-Guided Langevin Dynamics Simulation Method. *Chem. Phys. Lett.* **2003**, *381*, 512−518.
(39) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An *N*·Log(*N*) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(40) Christ, C. D.; Fox, T. Accuracy Assessment and Automation of Free Energy Calculations for Drug Design. *J. Chem. Inf. Model.* **2014**, *54*, 108−120.
(41) *GOLD 5.1*; Cambridge Crystallography Data Centre Software Ltd.: Cambridge, England, 2013.
(42) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Design, Synthesis, Biological Activity and Molecular Dynamics Studies of Specific Protein Tyrosine Phosphatase 1B Inhibitors over SHP-2. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.
(43) Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D. Flexible Docking Using Tabu Search and an Empirical Estimate of Binding Affinity. *Proteins* **1998**, *33*, 367−382.
(44) Lang, P. T.; Moustakas, D.; Brozell, S.; Carrascal, N.; Mukherjee, S.; Balius, T.; Allen, W. J.; Holden, P.; Pegg, S.; Raha, K.; Shivakumar, S.; Rizzo, R.; Case, D.; Shoichet, B.; Kuntz, I. *Dock 6.5*; University of California: Oakland, CA, 2013.
(45) Mikulskis, P.; Genheden, S.; Rydberg, P.; Sandberg, L.; Olsen, L.; Ryde, U. Binding Affinities of the Sampl3 Trypsin and Host-Guest Blind Tests Estimated with the MM/PBSA and LIE Methods. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 527−541.
(46) Brown, S. P.; Muchmore, S. W.; Hajduk, P. J. Healthy Skepticism: Assessing Realistic Model Performance. *Drug Discovery Today* **2009**, *14*, 420−427.
(47) Genheden, S.; Cabedo Martinez, A. I.; Criddle, M. P.; Essex, J. W. Extensive All-Atom Monte Carlo Sampling and QM/MM Corrections in the SAMPL4 Hydration Free Energy Challenge. *J. Comput. Aided Mol. Des.* **2014**, *28*, 187−200.
(48) Homeyer, N.; Stoll, F.; Hillisch, A.; Gohlke, H. Binding Free Energy Calculations for Lead Optimization: Assessment of Their Accuracy in an Industrial Drug Design Context. *J. Chem. Theory Comput.* **2014**, *10*, 3331−3344.
(49) Bhattacharyya, A. On a Measure of Divergence Between Two Statistical Populations Defined by Their Probability Distributions. *Bull. Cal. Math. Soc.* **1943**, *35*, 99−109.
(50) Wu, D.; Kofke, D. A. Phase-Space Overlap Measures. I. Fail-Safe Bias Detection in Free Energies Calculated by Molecular Simulation. *J. Chem. Phys.* **2005**, *123*, 054103.
(51) Pohorille, A.; Jarzynski, A.; Chipot, C. Good Practices in Free-Energy Calculations. *J. Chem. Phys. B* **2010**, *114*, 10235−10253.
(52) Li, H.; Yang, W. Forging The Missing Link in Free Energy Estimations: λ-WHAM in Thermodynamic Integration, Overlap Histogramming, and Free Energy Perturbation. *Chem. Phys. Lett.* **2007**, *440*, 155−159.
(53) Nichols, S. E.; Baron, R.; Ivetac, A.; McCammon, J. A. Predictive Power of Molecular Dynamics Receptor Structures in Virtual Screening. *J. Chem. Inf. Model.* **2011**, *51*, 1439−1446.
(54) Vedula, L. S.; Brannigan, G.; Economou, N. J.; Xi, J.; Hall, M. A.; Liu, R.; Rossi, M. J.; Dailey, W. P.; Grasty, K. C.; Klein, M. L.; Eckenhoff, R. G.; Loll, P. J. A Unitary Anesthetic Binding Site at High Resolution. *J. Biol. Chem.* **2009**, *284*, 24176−24184.
(55) Wang, S. D.; Meades, C.; Wood, G.; Osnowski, A.; Anderson, S.; Yuill, R.; Thomas, M.; Mezna, M.; Jackson, W.; Midgley, C.; Griffiths, G.; Fleming, I.; Green, S.; Mcnae, I.; Wu, S. Y.; Mcinnes, C.; Zheleva, D.; Walkinshaw, M. D.; Fischer, P. M. Synthesis and Biological Activity of 2-Anilino-4-(1H-Pyrrol-3-yl) Pyrimidine CDK Inhibitors. *J. Med. Chem.* **2004**, *47*, 1662−1675.
(56) Penning, T. D.; Talley, J. J.; Bertenshaw, S. R.; Carter, J. S.; Collins, P. W.; Docter, S.; Graneto, M. J.; Lee, L. F.; Malecha, J. W.; Miyashiro, J. M.; Rogers, R. S.; Rogier, D. J.; Yu, S. S.; Anderson, G. D.; Burton, E. G.; Cogburn, J. N.; Gregory, S. A.; Koboldt, C. M.; Perkins, W. E.; Seibert, K.; Veenhuizen, A. W.; Zhang, Y. Y.; Isakson, P. C. Synthesis and Biological Evaluation of the 1,5-Diarylpyrazole Class of Cyclooxygenase-2 Inhibitors: Identification of 4-[5-(4-Methylphenyl)-3- (trifluoromethyl)-1H-pyrazol-1-yl]-benzenesulfonamide (SC-58635, Celecoxib). *J. Med. Chem.* **1997**, *40*, 1347−1365.
(57) Yuthavong, Y.; Vilaivan, T.; Chareonsethakul, N.; Kamchonwongpaisan, S.; Sirawaraporn, W.; Quarrell, R.; Lowe, G. Combined Spatial Limitation around Residues 16 and 108 of Plasmodium falciparum Dihydrofolate Reductase Explains Resistance to Cycloguanil. *J. Med. Chem.* **2000**, *43*, 2738−2744.
(58) Tarnchompoo, B.; Sirichaiwat, C.; Phupong, W.; Intaraudom, C.; Sirawaraporn, W.; Kamchonwongpaisan, S.; Vanichtanankul, J.; Thebtaranonth, Y.; Yuthavong, Y. Development of 2,4-Diaminopyrimidines as Antimalarials Based on Inhibition of the S108N and C59R +S108N Mutants of Dihydrofolate Reductase from Pyrimethamine-Resistant Plasmodium Falciparum. *J. Med. Chem.* **2002**, *45*, 1244−1252.
(59) Kim, S.; Wu, J. Y.; Birzin, E. T.; Frisch, K.; Chan, W.; Pai, L.-Y.; Yang, Y. T.; Mosley, R. T.; Fitzgerald, P. M. D.; Sharma, N.; Dahllund, J.; Thorsell, A.-G.; DiNinno, F.; Rohrer, S. P.; Schaeffer, J. M.; Hammon, M. L. Estrogen Receptor Ligands. II. Discovery of Benzoxathiins as Potent, Selective Estrogen Receptor Alpha Modulators. *J. Med. Chem.* **2004**, *47*, 2171−2175.
(60) Matter, H.; Defossa, E.; Heinelt, U.; Blohm, P. M.; Schneider, D.; Müller, A.; Herok, S.; Schreuder, H.; Liesum, A.; Brachvogel, V.; Lönze, P.; Walser, A.; Al-Obeidi, F.; Wildgoose, P. Design and Quantitative Structure-Activity Relationship of 3-Amidinobenzyl-1h-Indole-2-Carboxamides as Potent, Nonchiral, and Selective Inhibitors of Blood Coagulation Factor Xa. *J. Med. Chem.* **2002**, *45*, 2749−2769.
(61) Watson, K. A.; Chrysina, E. D.; Tsitsanou, K. E.; Zographos, S. E.; Archontis, G.; Fleet, G. W. J.; Oikonomakos, N. G. Kinetic and Crystallographic Studies of Glucopyranose Spirohydantoin and Glucopyranosylamine Analogs Inhibitors of Glycogen Phosphorylase. *Proteins* **2005**, *61*, 966−983.
(62) Dreyer, G. B.; Lambert, D. M.; Meek, T. D.; Carr, T. J.; Tomaszek, T. A., Jr.; Fernandez, A. V.; Bartus, H.; Cacciavillani, E.; Hassel, A. M.; Minnich, M.; Petteway, S. R., Jr.; Metcalf, B. W. Hydroxyethylene Isostere Inhibitors of Human Immunodeficiency Virus-1 Protease: Structure-Activity Analysis Using Enzyme Kinetics, X-Ray Crystallography, and Infected T-Cell Assays. *Biochemistry* **1992**, *31*, 6646−6659.
(63) Smith, P. W.; Sollis, S. L.; Howes, P. D.; Cherry, P. C.; Starkey, I. D.; Cobley, K. N.; Weston, H.; Scicinski, J.; Merritt, A.; Whittington, A.; Wyatt, P.; Taylor, N.; Green, D.; Bethell, R.; Madar, S.; Fenton, R. J.; Morley, P. J.; Pateman, T.; Beresford, A. Dihydropyrancarboxamides Related to Zanamivir: A New Series of Inhibitors of Influenza Virus Sialidases. 1. Discovery, Synthesis, Biological Activity, and Structure−Activity Relationships of 4-Guanidino- and 4-Amino-4H-pyran-6-carboxamides. *J. Med. Chem.* **1998**, *41*, 787−797.
(64) Shirts, M. R.; Pitera, J. Q.; Swope, W. C.; Pande, V. S. Extremely Precise Free Energy Calculations of Amino Acid Side Chain Analogs: Comparison of Common Molecular Mechanics Force Fields for Proteins. *J. Chem. Phys.* **2003**, *119*, 5740−5761.

(65) 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*, 804108.
(66) Lawrenz, M.; Baron, R.; Wang, Y.; McCammon, J. A. Effects of Biomolecular Flexibility on Alchemical Calculations of Absolute Binding Free Energies. *J. Chem. Theory Comput.* **2011**, *7*, 2224−2232.
(67) Wu, X.; Brooks, B. R. Self-Guided Langevin Dynamics Simulation Method. *Chem. Phys. Lett.* **2003**, *381*, 512−518.
(68) Woods, C. J.; Essex, J. W.; King, M. A. The Development of Replica-Exchange-Based Free-Energy Methods. *J. Phys. Chem. B* **2003**, *107*, 13703.
(69) Hamelberg, D.; Mongan, J.; McCammon, J. A. Accelerated Molecular Dynamics: A Promising and Efficient Simulation Method for Biomolecules. *J. Chem. Phys.* **2004**, *120*, 11919.
(70) Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. J. Replica Exchange with Solute Tempering: A Method for Sampling Biological Systems in Explicit Water. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 13749−13754.
(71) Zheng, L.; Chen, M.; Yang, W. Random Walk in Orthogonal Space to Achieve Efficient Free-Energy Simulation of Complex Systems. *Proc. Natl. Acad. Sci.* **2008**, *105*, 20227−20232.
(72) 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.
(73) Zheng, L.; Yang, W. Practically Efficient and Robust Free Energy Calculations: Double-Integration Orthogonal Space Tempering. *J. Chem. Theory Comput.* **2012**, *8*, 810−823.
(74) Wang, L.; Friesner, R. A.; Berne, B. J. Replica Exchange with Solute Scaling: A More Efficient Version of Replica Exchange with Solute Tempering (REST2). *J. Phys. Chem. B* **2011**, *115*, 9431−9438.