


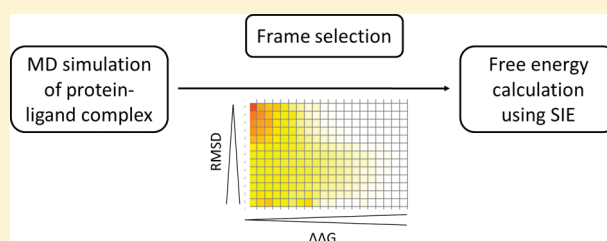
# Solvent Interaction Energy Calculations on Molecular Dynamics Trajectories: Increasing the Efficiency Using Systematic Frame Selection

Markus A. Lill\* and Jared J. Thompson

Department of Medicinal Chemistry and Molecular Pharmacology, College of Pharmacy, Purdue University, 575 Stadium Mall Drive, West Lafayette, Indiana 47907, United States

 Supporting Information

**ABSTRACT:** End-point methods such as linear interaction energy (LIE) analysis, molecular mechanics generalized Born solvent-accessible surface (MM/GBSA), and solvent interaction energy (SIE) analysis have become popular techniques to calculate the free energy associated with protein–ligand binding. Such methods typically use molecular dynamics (MD) simulations to generate an ensemble of protein structures that encompasses the bound and unbound states. The energy evaluation method (LIE, MM/GBSA, or SIE) is subsequently used to calculate the energy of each member of the ensemble, thus providing an estimate of the average free energy difference between the bound and unbound states. The workflow requiring both MD simulation and energy calculation for each frame and each trajectory proves to be computationally expensive. In an attempt to reduce the high computational cost associated with end-point methods, we study several methods by which frames may be intelligently selected from the MD simulation including clustering and address the question of how the number of selected frames influences the accuracy of the SIE calculations.



## INTRODUCTION

Docking is widely used in the drug discovery process for virtual screening and to predict protein–ligand binding modes. In order to dock large ligand libraries used in virtual screening, a balance between accuracy and efficiency must be found when modeling the underlying physics of protein–ligand interactions. Consequently, the scoring functions used to quantify protein–ligand interactions in docking programs typically focus on an important, but relatively simple, subset of interaction elements, such as hydrogen bonds and hydrophobic contacts. This simplified representation of protein–ligand interactions significantly contributes to the failure of docking to accurately predict binding affinities.<sup>1,2</sup> Although recent docking methods have been refined to somewhat include protein flexibility known to be important for protein–ligand conformational adaptation, the docking predicted free energy of binding is still usually based on a single protein–ligand complex structure, neglecting the fact that the protein–ligand complex samples local substates in the conformational vicinity of a given binding mode.

Postprocessing methods provide a means to overcome the weaknesses of simple scoring functions used in docking by including the missing dynamic information in the energy estimate and utilizing a more sophisticated physical representation of protein–ligand interaction, thus more accurately estimating the free energy of binding.<sup>3</sup> These techniques employ MD or MC simulations to provide a trajectory, and a free-energy estimation technique such as free-energy perturbation (FEP),<sup>4</sup> thermodynamic

integration (TI),<sup>5</sup> molecular-mechanics Poisson–Boltzmann, generalized-Born surface-area (MMPBSA/GBSA),<sup>6,7</sup> or linear interaction energy analysis (LIE)<sup>8</sup> is used to calculate the average energy over the trajectory. In docking applications that employ end-point methods (MMPBSA/GBSA, LIE, or SIE), either the top-scored binding pose or several favorably scored poses are used as input for the subsequent MD or MC simulations. Even if limited to a few ligands and a small number of binding poses, this postprocessing step can significantly improve the successful identification of binding modes and prediction of binding affinities.<sup>3</sup>

MMGBSA and LIE have become popular postprocessing methods due to their computational efficiency and applicability to diverse sets of ligands, drawbacks associated with the most accurate methods to predict relative free energies of binding, FEP and TI. LIE, however, requires a priori knowledge of a set of active ligands with experimentally known binding affinities in order to optimize the protein-dependent regression coefficients inherent to the LIE equations. In contrast, MMGBSA can be applied to any protein–ligand system without additional regression, but this method requires the calculation of an explicit entropy term that is prone to slow convergence<sup>9</sup> and, for some systems, displays overly large contributions to the absolute free energy of binding.<sup>10</sup> Other end-point methods used to quantify

**Received:** April 29, 2011

**Published:** August 28, 2011

**Table 1.** Protein–Ligand Complexes Used in Our Study: The Ligand Name (as used in this paper), the 2D Representation of Each Structure, the PDB Code of the Protein Structure of Each Complex, and the Binding Affinity of Each Ligand<sup>a</sup>

Ligand name	Ligand structure	PDB of protein	Affinity in nM	Ligand name	Ligand structure	PDB of protein	Affinity in nM
Neuraminidase							
N1		1bji	2	N2		1nnc	5
N3		1mwe	1·10 <sup>6</sup>	N4		2qwi	20
N5		2qwk	2	N6		1f8b	8600
N7		1f8c	320	N8		1bji	5
N9		1bji	320	N10		1bji	12000
Avidin							
A1		1avd	1.4·10 <sup>-6</sup>	A2		1avd	0.038
A3		1avd	0.063	A4		1avd	390
Avidin							
A5		1avd	1060	A6		1avd	2.3·10 <sup>5</sup>
A7		1avd	5.3·10 <sup>5</sup>				
Thrombin							
T1		1mu6	4.2	T2		1mu6	280
T3		1mu6	17000	T4		1mu6	0.042
T5		1mu6	44	T6		1mu6	0.5
T7		1mu6	12	T8		1mu6	0.36
T9		1mu6	2	T10		1mu6	3

<sup>a</sup> Experimental affinities are taken from the literature.<sup>25–32</sup>

protein–ligand interactions include the mining minima approach,<sup>11–14</sup> linear response approximation (LRA), and the protein dipoles Langevin dipoles (PDLD/S-LRA) version thereof.<sup>10,15,16</sup>

Solvated interaction energy (SIE)<sup>17</sup> is a relatively new end-point method that shares elements from the LIE and MMPBSA/GBSA methods. Similar to MMPBSA/GBSA, SIE treats the protein–ligand system in atomistic detail and solvation effects implicitly. The free energy of binding between ligand and protein is computed by:

$$\Delta G_{\text{bind}}(\rho, D_{\text{in}}, \alpha, \gamma, C) = \alpha[\Delta E_{\text{vdW}} + \Delta E_{\text{Coul}}(D_{\text{in}}) + \Delta G_{\text{RF}}(\rho, D_{\text{in}}) + \gamma \Delta SA(\rho)] + C \quad (1)$$

where  $\Delta E_{\text{vdW}}$  and  $\Delta E_{\text{Coul}}$  are the intermolecular van der Waals and Coulomb interaction energy between protein and ligand,  $\Delta G_{\text{RF}}(\rho, D_{\text{in}})$  is the difference in the reaction-field energy between the bound and free state of the protein–ligand complex as calculated by solving the Poisson equation with BRI BEM,<sup>18,19</sup> and  $\Delta SA(\rho)$  is the difference in molecular surface area between the bound and free state of the protein. The five parameters in eq 1 were fitted to the absolute free energy of binding of 99 protein–ligand complexes: The linear scaling factor  $\rho$  of the

van der Waals radii of the AMBER99 force field, the dielectric constant inside the solute  $D_{\text{in}}$ , the coefficient  $\gamma$  for quantifying the free energy associated with the difference in surface area upon protein–ligand binding, and the prefactor  $\alpha$  that implicitly quantifies the loss of entropy upon binding, also known as entropy–enthalpy compensation, and a constant  $C$  that includes protein-dependent contributions not explicitly modeled by the SIE methodology, e.g., the change in protein internal energy upon ligand binding. The default values of the parameters are:  $\rho = 1.1$ ,  $D_{\text{in}} = 2.25$ ,  $\gamma = 0.0129 \text{ kcal}/(\text{mol} \cdot \text{\AA}^2)$ ,  $C = -2.89 \text{ kcal}/\text{mol}$ , and  $\alpha = 0.1048$ .

SIE has been utilized to estimate the binding free energy based on a MD trajectory of the protein–ligand complex.<sup>20,21</sup> In this process, individual SIE calculations on equally separated snapshots from the trajectory are averaged to provide an estimate of the free energy of binding. However, studies seldom address the question of how many snapshots from the MD simulation are required to accurately predict the binding free energy. In this article we aim to address this question and focus on ways to reduce the computational time needed to accurately estimate binding energies using SIE. In particular, we address the following two questions: How does the number of snapshots used in the SIE calculation influence the accuracy of predicting the free energy of binding, and can we intelligently select frames from the

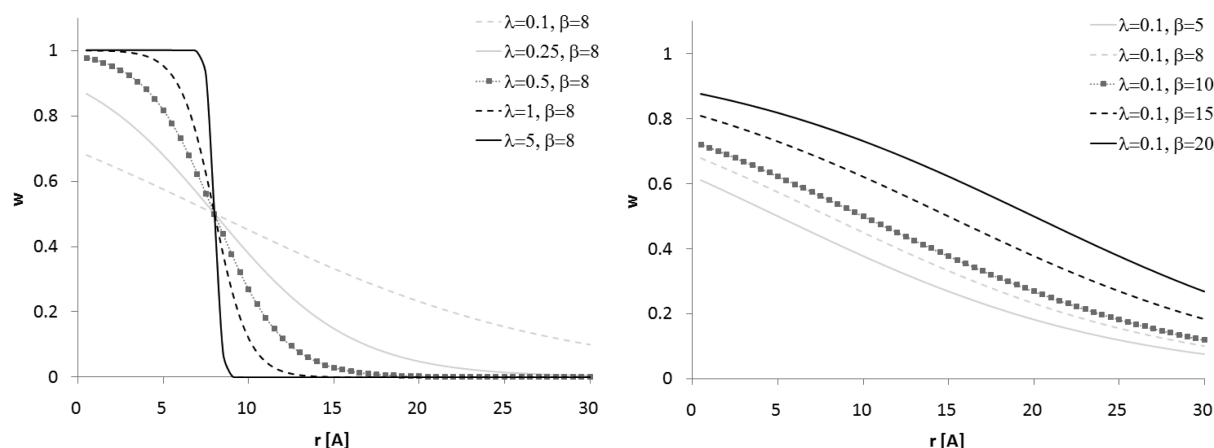


Figure 1. Examples of weighting functions used to cluster the MD frames.

Table 2. Mean Absolute Deviation between the Predicted and the Experimental Free Energies When Using the SIE (eq 1) Default Coefficients for  $\alpha$ ,  $\gamma$ , and  $C$  and Optimized Parameters for Each Protein System<sup>a</sup>

protein system	mean absolute deviation from $G_{\text{exp}}$ (no parameter optimization) (kcal/mol)	mean absolute deviation from $G_{\text{exp}}$ (parameter optimization) (kcal/mol)	optimized parameters		
			$\alpha$	$\gamma$ [kcal/(mol·Å <sup>2</sup> )]	$C$ (kcal/mol)
neuraminidase	2.26	1.39	0.22	0.023	2.5
avidin	2.71	1.78	0.34	0.005	2.5
thrombin	2.86	1.47	0.48	0.024	20.5

<sup>a</sup> The default values are  $\alpha = 0.1048$ ,  $\gamma = 0.0129$  kcal/(mol·Å<sup>2</sup>), and  $C = -2.89$  kcal/mol.

MD simulation that represent structurally similar frames with similar contributions to the binding energy by clustering the full trajectory? This article can be related to other work studying the convergence of alternative end-point methods such as MMPBSA and MMGBSA.<sup>22–24</sup>

## MATERIALS AND METHODS

**Protein Systems and Preparation.** Our study was performed on three different protein systems: neuraminidase, avidin, and thrombin. For neuraminidase, ten protein–ligand complexes were studied containing seven experimentally determined crystal structures (1bji, 1nnc, 1mwe, 2qwi, 2qwk, 1f8c, 1f8b) and three additional complexes by adding three ligands (Table 1, N8–N10) to the 1bji structure.<sup>25</sup> For these three complexes, the initial binding pose of the original 5-acetylamino-4-amino-6-(phenethylpropylcarbamoyl)-5,6-dihydro-4H-pyran-2-carboxylic acid ligand was used, but the propyl group was shortened to an ethyl group, a methyl group, or a hydrogen atom to generate the three additional pseudo X-ray structures (Table 1, N8–N10). For avidin, seven ligands were chosen that were previously used in MM/PBSA<sup>26</sup> and LIE<sup>27</sup> studies. Based on the biotin–avidin complex (1avd), six additional ligands (Table 1, A2–A7) were generated by manual mutation of the biotin ligand in the binding site of avidin. For thrombin, we used a data set containing ten ligands from a single SAR study<sup>28–32</sup> and manually mutated the cocrystallized ligand from the 1mu6 crystal structure to generate the starting complex structures of thrombin with ligands T1–T10.

All ligands and their associated binding affinities are displayed in Table 1.

The hydrogen-bond network of each crystal structure was optimized by rotating Asn, Gln, and His residues and by assigning protonation states to His using the Reduce program.<sup>33</sup> The protein parameters used in the molecular mechanics minimizations and molecular dynamics (MD) simulations were assigned based on the Amber ff03<sup>34</sup> force field implemented in the Amber10 program suite.<sup>35</sup> The ligand force field parameters were assigned using the general Amber force field (gaff),<sup>36</sup> and partial charges were calculated using the AM1/BCC methodology.<sup>37</sup> Each protein–ligand system was placed in a rectangular box of TIP3P water with the minimum distance between any solute atom and the boundary of the box set to 10 Å. The system was neutralized by adding Cl<sup>−</sup> or Na<sup>+</sup> counterions as needed. All protein–ligand systems were prepared using our in-house PyMOL<sup>38</sup> plugin that automatically calls the antechamber and tleap modules from the AmberTools 1.4 program suite.

**MD Simulation Protocol.** Periodic boundary conditions were applied to each protein–ligand system and the long-range electrostatic interactions were calculated using the particle mesh Ewald (PME) method. After 500 steps of minimization, the water molecules were equilibrated for 250 ps restraining the position of any solute atom with an harmonic potential with a force constant of 5 kcal/(mol·Å<sup>2</sup>). Subsequently, the full system was equilibrated for 500 ps, and the final production MD simulation was run for 10 ns for each system. The time step was 2 fs, the temperature set to 310 K, and all bonds containing a hydrogen atom were constrained using the SHAKE algorithm.<sup>39</sup>

**Table 3. The Accuracy of Predicting Binding Free Energies Using SIE As Measured by the Pearson Correlation Coefficient ( $r$ ) between Experimental and Predicted Binding Free Energies and the Leave-One-Out Cross-Validated Pearson Coefficient ( $q$ )<sup>a</sup>**

protein system	Pearson correlation coefficient for full MD trajectory	leave-one-out cross-validated Pearson coefficient for full MD trajectory
neuraminidase	0.83	0.69
avidin	0.89	0.77
thrombin	0.71	0.41

<sup>a</sup> The  $r$  and  $q$  values are shown for three different protein systems using all 1000 snapshots from each MD simulation.

All molecular mechanics minimizations and position restraint simulations were performed with the sander module of the Amber10 program suite, and the equilibration and production runs used the pmemd module. One thousand equally spaced snapshots of the protein–ligand complex (every 10 ps) were generated from the production MD trajectory, and all water molecules and counterions were removed before subsequent SIE analysis. This ensemble of 1000 snapshots is referred to as “full trajectory” throughout the article.

**Free Energy Calculation Using SIE.** The free energy of ligand binding was determined by applying the sietraj<sup>17</sup> software to the selected MD snapshots and averaging over the resulting free energies obtained from each snapshot. The correlation between the predicted and experimental free energy was determined using the default parameters in eq 1 as well as optimized parameters for  $\alpha$ ,  $\gamma$ , and  $C$ . To obtain optimized parameters for each protein system, the parameters  $\alpha$  and  $\gamma$  were systematically varied within physically meaningful ranges ( $\alpha \in [0.05; 1.0]$ ,  $\gamma \in [0.005; 0.025]$  kcal/(mol·Å<sup>2</sup>)), and  $C$  was optimized to minimize the sum of the absolute deviations between predicted and experimental affinity for all ligands in a protein data set. The values for  $\alpha$  need to be positive and smaller than one as they characterize the entropy–enthalpy compensation, and  $\gamma$  should be in a range postulated by other studies<sup>40,41</sup> utilizing the difference in molecular surface area as the contribution to the free energy of binding associated with nonpolar desolvation.

**Selection of Snapshots.** SIE calculations were performed using sietraj<sup>17</sup> on the full MD trajectory and on selected frames from the MD simulation. The following frame selection protocols were investigated: Equally spaced snapshots from each MD trajectory were selected using 1, 2, 3, 4, 5, 7, 10, 15, 20, 25, and 50 frames for separate analyses. Random selection of frames from the full trajectory was performed five times, yielding five different analyses for each number of frames (1, 2, 3, 4, 5, 7, 10, 15, 20, 25, 50) to be selected. As a last selection protocol, the MD trajectory was clustered based on pairwise rmsd values between different snapshots from the MD trajectory. The rmsd values between different MD frames  $s$  and  $t$  were computed using only the heavy atoms of the protein:

$$\text{RMDS}(s, t) = \frac{\sum_{i \in \text{heavyatoms}} w_i \sqrt{|r_i(s) - r_i(t)|^2}}{\sum_{i \in \text{heavyatoms}} w_i} \quad (2)$$

and the weight  $w$  of each protein atom's contribution to the rmsd was calculated using the formula:

$$w_i = \frac{1}{\exp(\lambda(\tilde{r}_{iL} - \beta)) + 1} \quad (3)$$

where  $\tilde{r}_{iL}$  is the shortest distance between protein atom  $i$  and any heavy ligand atom in the initial frame of the MD simulation, and  $\lambda$  and  $\beta$  characterize the size and slope of the weight as a function of distance (see Figure 1). Thus, we weight protein atoms closer to the ligand more than atoms distant from it, in order to bias the calculation of rmsd toward the region surrounding the ligand. Different sets of values for parameters  $\lambda$  (0.1 Å<sup>−1</sup>, 0.25 Å<sup>−1</sup>, 0.5 Å<sup>−1</sup>, 1.0 Å<sup>−1</sup>, 5.0 Å<sup>−1</sup>) and  $\beta$  (5 Å, 8 Å, 10 Å, 15 Å, and 20 Å) were investigated (Figure 1). It should be noted that curves with  $\lambda = 5.0$  Å<sup>−1</sup> approximate the standard rmsd calculations but include a cutoff defined by the parameter  $\beta$  (see Figure 1, left).

K-means clustering was performed using the rmsd values to generate 1, 2, 3, 4, 5, 7, 10, 15, 20, 25, and 50 clusters from the MD trajectory. The frame with the smallest sum of rmsd values to all members of the cluster is selected as the representative protein–ligand structure and subsequently used as the input for SIE calculations.

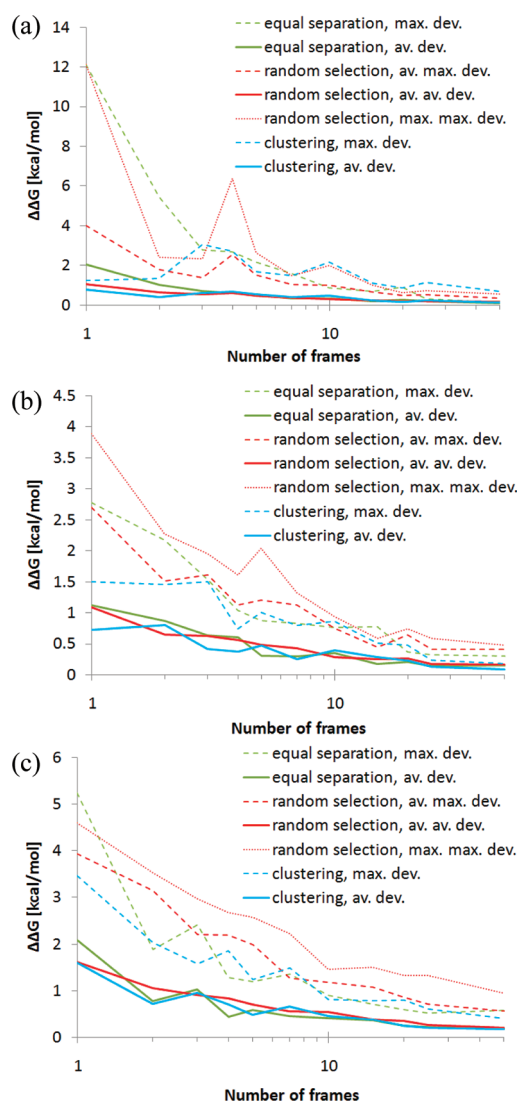
**Measures of Prediction Accuracy.** Two different criteria were selected to measure the influence of frame selection on the prediction quality of the free energies of binding. The first measure assumes that the SIE calculation on the full MD trajectory (containing 1000 frames) is the most precise estimation of the free energy of binding, referred to as “full” predicted free energy. The difference between the free energy computed using a reduced set of MD frames and the predicted free energy using the full MD trajectory serves as the criterion to define the accuracy of computing the free energy of binding. The second measure of accuracy is the difference between the Pearson regression coefficient ( $r$ ) of the free energy of binding calculated when using the full trajectory and the reduced number of snapshots. A comparable analysis was performed using the Spearman's rank correlation coefficient ( $r_s$ ). The results for this analysis are shown in the Supporting Information S1.

## RESULTS AND DISCUSSION

### Prediction Accuracy of SIE Using the Full MD Trajectory.

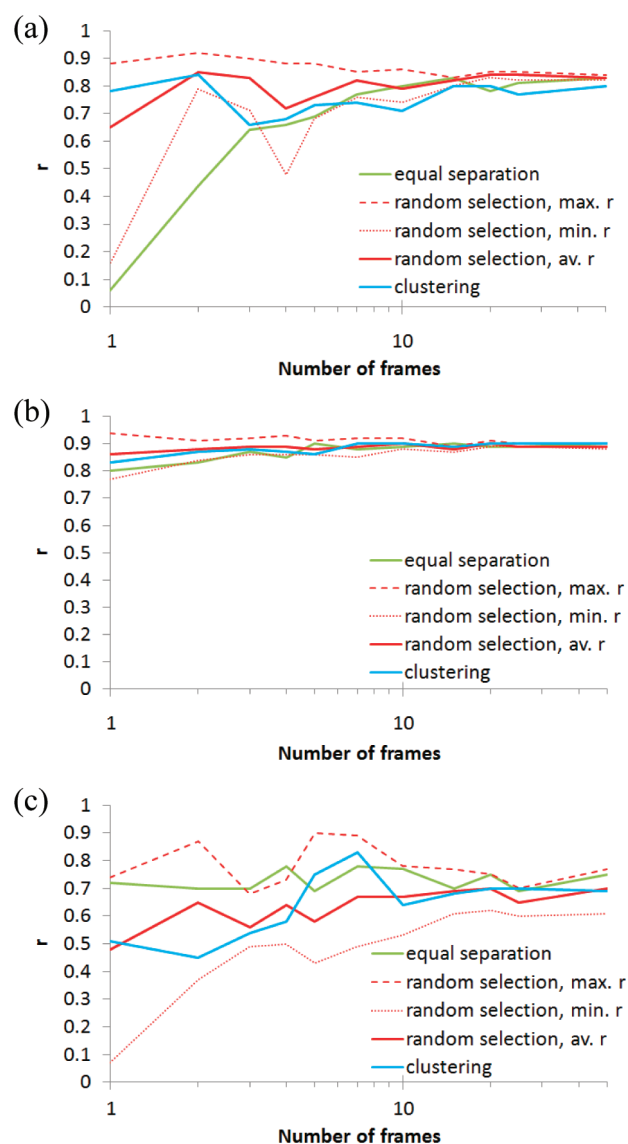
To establish an “upper bound” for SIE's ability to calculate binding free energies for the three selected protein systems, we computed the “full” predicted binding free energy based on all snapshots of the MD trajectories. Comparing the “full” SIE predicted free energies with experimental free energies ( $G_{\text{exp}}$ ) revealed significant deviations (see Table 2) when using SIE's default set of parameters (optimized on 99 PDB structures). In an attempt to improve the prediction of the absolute binding free energies, we optimized the SIE parameters  $\alpha$ ,  $\gamma$ , and  $C$ . The predicted binding free energies using the optimized parameter set for each protein system reveals a significant reduction in the mean absolute deviation between the predicted and the experimental free energies of binding (Table 2). All three parameters ( $\alpha$ ,  $\gamma$ , and  $C$ ) were significantly different than the default value, and different optimized parameter values were identified for each individual protein system. A 2-fold decrease (avidin) and increase (neuraminidase, thrombin) was observed for  $\gamma$  compared to the default value, and the parameter  $\alpha$  increased by approximately a factor 2–5 in all protein systems. The largest deviation between default and fitted value was observed for the parameter





**Figure 2.** Absolute deviation of the free energy predicted from a set number of frames extracted from the MD trajectory compared to the free energy computed on the full MD trajectory. Displayed are the deviations as a function of number of selected frames (logarithmic scale) for the protein systems (a) neuraminidase, (b) avidin, and (c) thrombin. The results of using equally spaced frames (green line: absolute deviation averaged over all ligands; green dashed line: maximum absolute deviation of an individual ligand), snapshots chosen by random selection (red line: absolute deviation averaged over all ligands and all five different random selection runs; red dashed line: maximum absolute deviation of all ligands averaged over all five different random selection runs; red dotted line: overall maximum absolute deviation of an individual ligand among any of the five different random selection runs), and frames identified by *k*-means clustering using the pairwise rmsd values between MD frames as distance criterion (blue line: absolute deviation averaged over all ligands; blue dashed line: maximum absolute deviation of an individual ligand) are shown. For comparison, the mean absolute deviation of the full trajectory from  $G_{\text{exp}}$  for neuraminidase, avidin, and thrombin are 1.39, 1.78, and 1.47 kcal/mol, respectively. Analysis of the individual contributions to the SIE energy (see Supporting Information S3) displays that the van der Waals ( $\alpha\Delta E_{\text{vdW}}$ ), electrostatic ( $\alpha\Delta E_{\text{Coul}}(D_{\text{in}})$ ), and reaction field ( $\alpha\Delta G_{\text{RF}}(D_{\text{in}})$ ) energy contributes about equally to the average deviation from the “full” free energy whereas the solvent accessible energy term ( $\alpha\gamma\Delta S_{\text{A}}(\rho)$ ) contributes less to the average deviation.

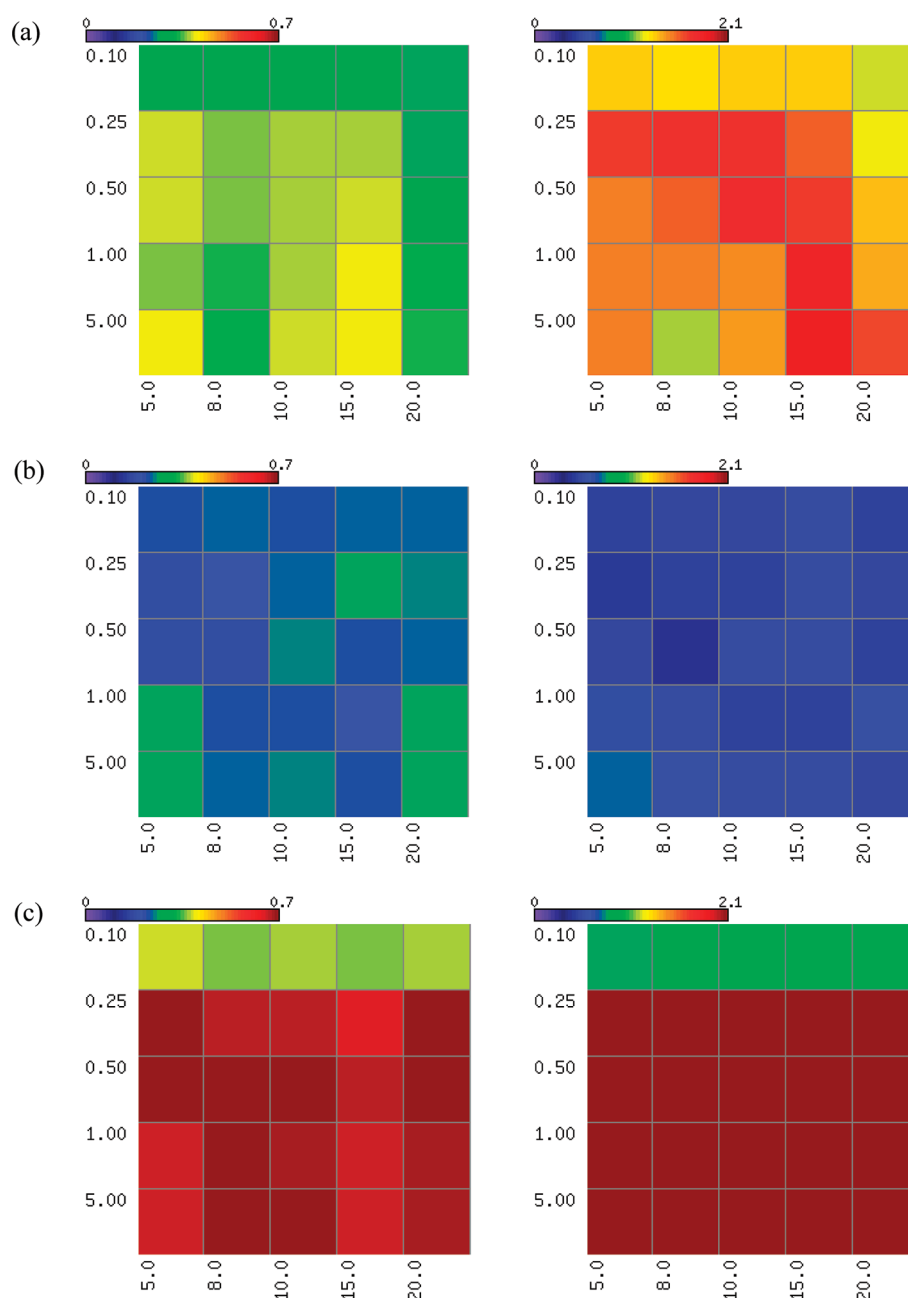
C. Potential reasons for the need to refit the three parameters for each system include the inherent shortcomings of the molecular-



**Figure 3.** Pearson regression coefficient (*r*) of the predicted free energies based on a selected number of frames compared to the free energy computed using the full MD trajectory. Displayed are the *r* values as a function of number of selected frames (logarithmic scale) for the protein systems (a) neuraminidase, (b) avidin, and (c) thrombin. The results of using equally spaced frames (green line), snapshots chosen by random selection (red dashed line: maximum *r* among the five random selection runs; red dotted line: the minimum *r* among the five random selection runs; red line: *r* value averaged over the five random selection runs), and frames identified by *k*-means clustering using the pairwise rmsd values between the MD frames as distance criterion (blue line).

mechanics force field (e.g., neglect of polarization, electron transfer, or lone-pair directionality), the neglect of internal energy changes using a single trajectory approach (i.e., no individual protein or ligand simulation is performed), and explicit solvent effects. Furthermore, the ratio of entropy to enthalpy is not necessarily identical for each protein–ligand complex. Finally, the experimental binding affinities themselves are not consistently measured for all three protein targets. A relative shift of binding free energies between the three protein systems is not unlikely.

Representative plots displaying SIE energy versus time and all-heavy atoms rmsd (to the original protein–ligand complex



**Figure 4.** The fitness function  $f$  (eq 4) dependent on variables  $\lambda$  (vertical axis; units in  $\text{\AA}^{-1}$ ) and  $\beta$  (horizontal axis; units in  $\text{\AA}$ ) is shown for (a) neuraminidase, (b) avidin, and (c) thrombin. An optimal combination of  $\lambda$  and  $\beta$  corresponds to the lowest fitness function value. The average deviations over all compounds of a protein system are displayed in the left column and the maximum individual deviation of an individual ligand in the right column.

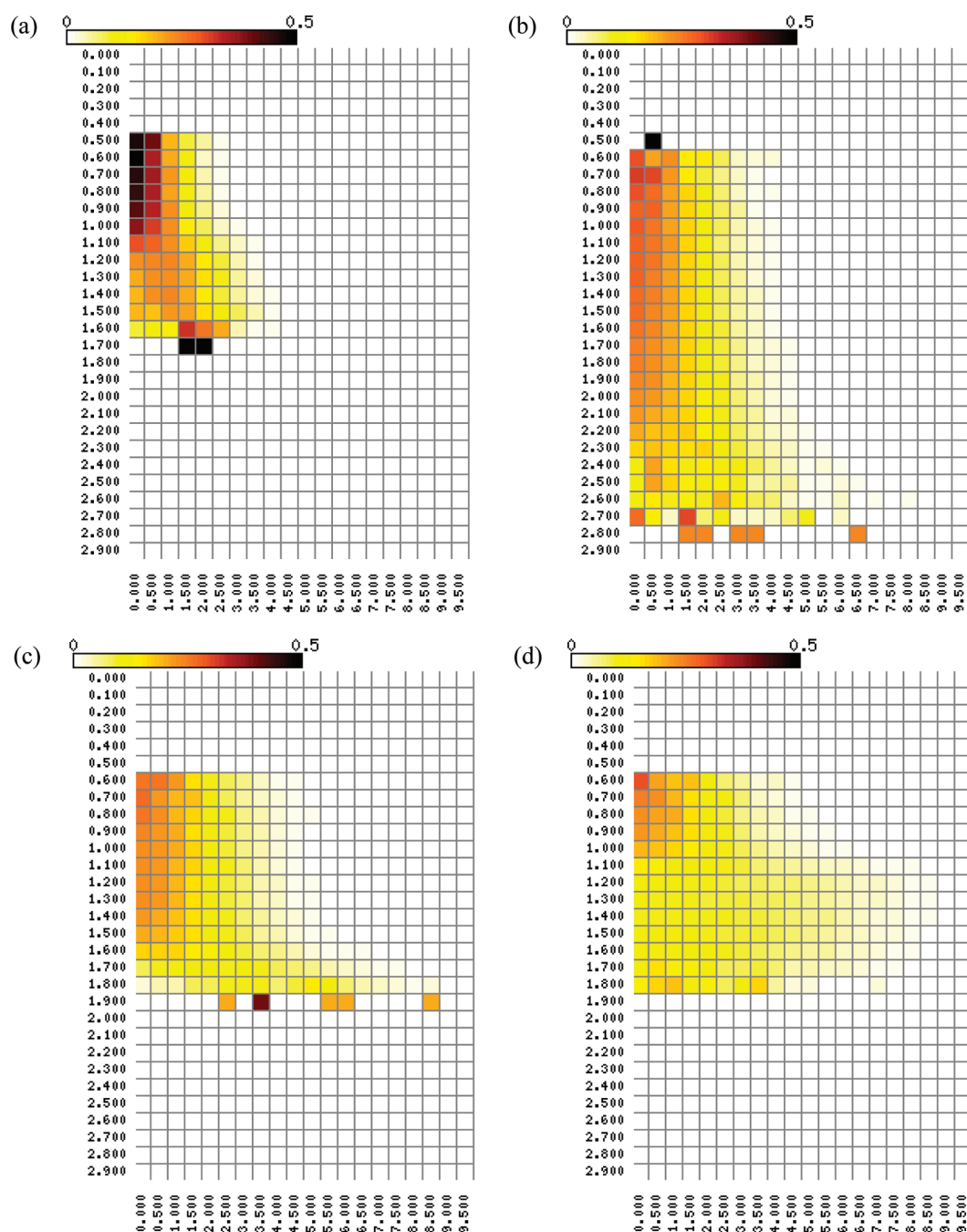
structure) versus time for each protein system are displayed in Supporting Information S2 indicating how such terms fluctuate over the “full” trajectory for each protein system.

Using the optimized set of SIE parameters (Table 2), the Pearson correlation coefficients ( $r$ ) between the experimental and SIE predicted free energy of binding were at least 0.7 for all three of the protein systems studied (shown in Table 3). To assess the generality of the regression model, we performed leave-one-out (LOO) cross-validation. The difference between the LOO Pearson correlation coefficient ( $q$ ) and  $r$  (see Table 3) suggests that the regression models for neuraminidase and avidin are relatively stable whereas a more significant difference between

$q$  and  $r$  was identified for thrombin. This is consistent with the largest deviation of the optimized parameters  $\alpha$ ,  $\gamma$ , and  $C$  from the default values for thrombin compared to neuraminidase and avidin.

#### Prediction Accuracy Based on a Selection of Snapshots.

Next, we address the question of how the number of frames selected from the MD simulation influences the accuracy of SIE calculated binding affinities. Using a selection of equally spaced frames from the MD trajectory, the absolute free energy difference between the SIE computed energy averaged over the selected MD snapshots and the mean energy using the full MD trajectory was computed. This deviation averaged over all complexes of each studied protein system quickly diminished to



**Figure 5.** Difference in the predicted free energy (horizontal axis; units in kcal/mol) between two MD snapshots as a function of the rmsd between the snapshots (vertical axis; units in Å) for four selected protein–ligand complexes: (a) compound N1 binding to 1bji (neuraminidase), (b) compound A3 binding to 1avd (avidin), (c) compound T4 and (d) compound T8 binding to 1mu6 (thrombin). The rmsd values are distributed into bins with a width of 0.1 Å, and the differences in the SIE energy are distributed to bins with a width of 0.5 kcal/mol. For each rmsd bin (row), the probability is normalized to one, and the color coding represents the probability of identifying a pair of MD frames with specified energy difference at given rmsd value. No pairs of frames with rmsd <0.5 Å were sampled throughout the individual MD simulations, and these lines are assigned a probability of zero. The same assignment is performed for large rmsd bins that were not sampled by the MD simulation. The magnitude of the probability is color coded, ranging from white (zero probability) over yellow and red to black (maximum probability).

approximately 0.5 kcal/mol as the number of selected frames (Figure 2, green line) increased. Furthermore, only five snapshots from each protein system were needed to reach this level of accuracy. The maximum absolute deviation for any compound in each protein data set is below 1 kcal/mol if ten equally separated MD frames (Figure 2, green dashed line) were used.

As an alternative to equally spaced frame selection, MD snapshots were randomly selected from the ensemble of 1000 frames from the full MD trajectory. For each amount of randomly selected frames (1, 2, 3, 4, 5, 7, 10, 15, 20, 25, 50), the random selection procedure was repeated five times. The mean (Figure 2, red line) and maximum (Figure 2, red dashed line) deviation

between the predicted and the experimental binding free energy, averaged over the five different random selection procedures, was comparable in size to the deviations observed for equally spaced frame selection. Considering each individual random selection run separately (i.e., one of the five runs), however, produced less rapid approximations (compared to equal separation) of the free energy of binding when increasing the number of frames. This is evident from the dotted red line in Figure 2, showing the maximum deviation between randomly selected frames and using the full trajectory to predict the free energy of binding for an individual ligand. The reduction in predictive accuracy when randomly selecting frames compared to selecting equally separated frames is intuitive, as fluctuations of the computed free energies over a 10 ns MD trajectory may be partially smoothed out by selecting equally separated frames, whereas random selection can choose frames that are chronologically close on the timeline of the MD simulation. When this situation occurs in random selection, close snapshots can overweight the free energy contribution of frames that display large deviation from the average SIE energy of the MD trajectory.

A similar relationship between the frame selection process and the prediction quality was observed when analyzing the variation of the Pearson correlation coefficient ( $r$ ) as a function of the number of selected frames (Figure 3). Using a selection of five equally separated frames,  $r$  deviated by less than 0.1 from the full trajectory correlation coefficient in all three protein systems. On average, a random selection of five frames produced the same level of prediction quality when the same number of equally spaced frames was used. It should be noted that  $r$  did not significantly increase with number of selected frames (cf. Figure 3) in all protein systems.

In an attempt to further improve the frame-selection process, snapshots were extracted from the full MD trajectory by clustering the trajectory using the pairwise rmsd values between the snapshots. The rmsd values between different snapshots were calculated using eq 2 and 3 with different  $\lambda$  and  $\beta$  values. To identify the optimal combination of the  $\lambda$  and  $\beta$  value, the average and maximum deviation (over all complexes of a protein system) of the predicted free energy of binding using  $k$  number of extracted frames (= cluster centers) from the “full” predicted free energy was calculated for  $k = 1, 2, 3, 4, 5, 7, 10, 15, 20, 25$ , and 50. The best combination was then selected by a fitness function that averages over the deviations  $d_k$  of particular number  $k$  weighted by the number of clustered snapshots  $k$ :

$$f = \frac{\sum_k k d_k}{\sum_k k} \quad \text{with } k \in \{1, 2, 3, 4, 5, 7, 10, 15, 20, 25, 50\} \quad (4)$$

The underlying idea of the fitness function is that predicted free energies based on a large number of frames are expected to have smaller deviations from the “full” free energy compared to energies that are based on a small number of frames. Consequently, large deviations using a large number of frames should be weighted more than similar sized deviations when using a smaller number of frames. The optimal values of  $\lambda$  and  $\beta$  are achieved when the fitness function is minimal.

Figure 4 displays the fitness function in form of a heat map for the average deviation over all compounds of a protein system (left column) and the maximum individual deviation of an individual ligand (right column) of the data set when the values of  $\lambda$  and  $\beta$  are varied. Evident in Figure 4, different combinations

of  $\lambda$  and  $\beta$  are optimal for each protein system, but a  $\lambda$  value of  $0.1 \text{ \AA}^{-1}$  consistently gives low fitness values for all three protein systems. Because no clear trend was identified when trying to determine the optimal value of  $\beta$  when  $\lambda = 0.1 \text{ \AA}^{-1}$ , we arbitrarily picked  $\beta = 8 \text{ \AA}$  for subsequent analysis and discussion.

The reasoning behind selecting frames by representative clusters of the MD trajectory is the assumption that similar protein–ligand structures results in similar protein–ligand interactions that are subsequently reflected in similar SIE energies. To test this hypothesis, we computed the rmsd values using eq 2 between all 1000 snapshots of the MD trajectory for randomly selected ligands from our data set and correlated the rmsd with the computed difference in SIE energies (Figure 5). The plots indeed support our hypothesis: low rmsd frame pairs tend to have small calculated energy differences, whereas frames with a large rmsd value between them frequently had large energy differences. However, the observed trend between rmsd and the energy difference between frame pairs is relatively weak and protein-system dependent. We note that for frame pairs with a low rmsd ( $<1 \text{ \AA}$ ), energy differences of up to 5 kcal/mol were observed. Due to this weak trend between structural and energetic similarity among the MD frames, the quality of the SIE-predicted binding free energy did not significantly improve using our trajectory clustering scheme compared to using the equally spaced frame selection scheme (Figures 2 and 3).

## CONCLUSIONS

Estimating free energies of binding using SIE is a potentially valuable addition to the tool set of end-point methods such as MM/GBSA and LIE. Studying three different protein systems, however, revealed that the default set of SIE parameters is not sufficient to accurately predict the absolute free energy of binding. Our studies suggest that retuning the parameter set, analogous to the standard tuning procedure used for LIE calculations, is necessary to achieve accurate calculations using SIE.

Remarkably, selecting 5–10 equally spaced snapshots displayed prediction accuracy comparable in quality to using the full MD trajectory. An analogy can be drawn to a study by Wang et al.<sup>42</sup> on consensus scoring, if we assume that the energies associated with the different frames fluctuate around the mean following a normal distribution, and that the mean corresponds to the experimental free energy. Wang et al.<sup>42</sup> showed that under this premise, the mean of about five energy values from different scoring functions approach the true value, in their study of the experimental binding affinity. On the basis of the idea that clustering of the MD trajectory would identify frames that cover the conformational space sampled by the protein–ligand complex, we utilized  $k$ -means clustering of the MD trajectory to select frames for SIE analysis. Because of the relatively minor trend between structural and energetic similarity between the MD frames when using SIE, the clustering scheme did not lead to an increase in prediction accuracy compared to using equally spaced frames from the trajectory.

On the basis of the reported studies, SIE appears to have the potential to be a powerful postprocessing tool that could be used to provide a better estimation of binding affinity for docking poses. It should, however, be noted that in this study we have used known binding poses and have not tested if SIE could differentiate native from decoy docking poses or can accurately predict binding affinities including structurally even more diverse families of ligands. In addition, only a small number of selected



frames are needed to achieve the same level of accuracy when using the entire MD trajectory to estimate the free energy of binding.

## ■ ASSOCIATED CONTENT

**S Supporting Information.** The difference between the Spearman's rank regression coefficient of the free energy of binding calculated when using the full trajectory and the reduced number of snapshots, representative plots displaying SIE energy versus time and all-heavy atoms rmsd versus time, and analysis of the individual contributions to the SIE energy. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## ■ AUTHOR INFORMATION

### Corresponding Author

\*E-mail: [mlill@purdue.edu](mailto:mlill@purdue.edu).

## ■ ACKNOWLEDGMENT

We thank Matthew Danielson for critical reading of the manuscript. This work has in part been supported by the National Institutes of Health (GM085604 and GM092855).

## ■ REFERENCES

- (1) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.
- (2) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L., III. Assessing scoring functions for protein–ligand interactions. *J. Med. Chem.* **2004**, *47*, 3032–3047.
- (3) Alonso, H.; Bliznyuk, A. A.; Gready, J. E. Combining docking and molecular dynamic simulations in drug design. *Med. Res. Rev.* **2006**, *26*, 531–568.
- (4) Zwanig, R. W. High-Temperature Equation of State by a Perturbation Method. 1. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22*, 1420–1426.
- (5) Kirkwood, J. G. Statistical mechanics of fluid mixtures. *J. Chem. Phys.* **1935**, *3*, 300–313.
- (6) Srinivasan, J.; Miller, J.; Kollman, P. A.; Case, D. A. Continuum solvent studies of the stability of RNA hairpin loops and helices. *J. Biomol. Struct. Dyn.* **1998**, *16*, 671–682.
- (7) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E., III. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000**, *33*, 889–897.
- (8) Aqvist, J.; Medina, C.; Samuelsson, J. E. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.* **1994**, *7*, 385–391.
- (9) Kongsted, J.; Ryde, U. An improved method to predict the entropy term with the MM/PBSA approach. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 63–71.
- (10) Singh, N.; Warshel, A. Absolute binding free energy calculations: on the accuracy of computational scoring of protein–ligand interactions. *Proteins* **2010**, *78*, 1705–1723.
- (11) Chang, C. E.; Gilson, M. K. Free energy, entropy, and induced fit in host–guest recognition: calculations with the second-generation mining minima algorithm. *J. Am. Chem. Soc.* **2004**, *126*, 13156–13164.
- (12) Chen, W.; Gilson, M. K.; Webb, S. P.; Potter, M. J. Modeling Protein–Ligand Binding by Mining Minima. *J. Chem. Theory Comput.* **2010**, *6*, 3540–3557.
- (13) David, L.; Luo, R.; Gilson, M. K. Ligand–receptor docking with the Mining Minima optimizer. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 157–171.
- (14) Kairys, V.; Gilson, M. K. Enhanced docking with the mining minima optimizer: acceleration and side-chain flexibility. *J. Comput. Chem.* **2002**, *23*, 1656–1670.
- (15) Lee, F. S.; Chu, Z. T.; Bolger, M. B.; Warshel, A. Calculations of antibody–antigen interactions: microscopic and semi-microscopic evaluation of the free energies of binding of phosphorylcholine analogs to McPC603. *Protein Eng.* **1992**, *5*, 215–228.
- (16) Sham, Y. Y.; Chu, Z. T.; Tao, H.; Warshel, A. Examining methods for calculations of binding free energies: LRA, LIE, PDL-DLRA, and PDL-DLRA calculations of ligands binding to an HIV protease. *Proteins* **2000**, *39*, 393–407.
- (17) Naim, M.; Bhat, S.; Rankin, K. N.; Dennis, S.; Chowdhury, S. F.; Siddiqi, I.; Drabik, P.; Sulea, T.; Bayly, C. I.; Jakalian, A.; Purisima, E. O. Solvated interaction energy (SIE) for scoring protein–ligand binding affinities. 1. Exploring the parameter space. *J. Chem. Inf. Model.* **2007**, *47*, 122–133.
- (18) Purisima, E. O.; Nilar, S. H. A Simple Yet Accurate Boundary-Element Method for Continuum Dielectric Calculations. *J. Comput. Chem.* **1995**, *16*, 681–689.
- (19) Purisima, E. O. Fast summation boundary element method for calculating solvation free energies of macromolecules. *J. Comput. Chem.* **1998**, *19*, 1494–1504.
- (20) Yang, B.; Hamza, A.; Chen, G. J.; Wang, Y.; Zhan, C. G. Computational Determination of Binding Structures and Free Energies of Phosphodiesterase-2 with Benzo[1,4]diazepin-2-one Derivatives. *J. Phys. Chem. B* **2010**, *114*, 16020–16028.
- (21) Wang, Y. T.; Su, Z. Y.; Hsieh, C. H.; Chen, C. L. Predictions of Binding for Dopamine D2 Receptor Antagonists by the SIE Method. *J. Chem. Inf. Model.* **2009**, *49*, 2369–2375.
- (22) Li, Y.; Liu, Z.; Wang, R. Test MM-PB/SA on true conformational ensembles of protein–ligand complexes. *J. Chem. Inf. Model.* **2010**, *50*, 1682–1692.
- (23) Genheden, S.; Ryde, U. How to obtain statistically converged MM/GBSA results. *J. Comput. Chem.* **2010**, *31*, 837–846.
- (24) Brown, S. P.; Muchmore, S. W. Rapid estimation of relative protein–ligand binding affinities using a high-throughput version of MM-PBSA. *J. Chem. Inf. Model.* **2007**, *47*, 1493–1503.
- (25) Taylor, N. R.; Cleasby, A.; Singh, O.; Skarzynski, T.; Wonacott, A. J.; Smith, P. W.; Sollis, S. L.; Howes, P. D.; Cherry, P. C.; Bethell, R.; Colman, P.; Varghese, J. Dihydropyranocarboxamides related to zanamivir: a new series of inhibitors of influenza virus sialidases. 2. Crystallographic and molecular modeling study of complexes of 4-amino-4H-pyran-6-carboxamides and sialidase from influenza virus types A and B. *J. Med. Chem.* **1998**, *41*, 798–807.
- (26) Kuhn, B.; Kollman, P. A. Binding of a diverse set of ligands to avidin and streptavidin: An accurate quantitative prediction of their relative affinities by a combination of molecular mechanics and continuum solvent models. *J. Med. Chem.* **2000**, *43*, 3786–3791.
- (27) Wang, J.; Dixon, R.; Kollman, P. A. Ranking ligand binding affinities with avidin: A molecular dynamics-based interaction energy study. *Proteins* **1999**, *34*, 69–81.
- (28) Burgey, C. S.; Robinson, K. A.; Lyle, T. A.; Sanderson, P. E. J.; Lewis, S. D.; Lucas, B. J.; Krueger, J. A.; Singh, R.; Miller-Stein, C.; White, R. B.; Wong, B.; Lyle, E. A.; Williams, P. D.; Coburn, C. A.; Dorsey, B. D.; Barrow, J. C.; Stranieri, M. T.; Holahan, M. A.; Sitko, G. R.; Cook, J. J.; McMasters, D. R.; McDonough, C. M.; Sanders, W. M.; Wallace, A. A.; Clayton, F. C.; Bohn, D.; Leonard, Y. M.; Detwiler, T. J.; Lynch, J. J.; Yan, Y. W.; Chen, Z. G.; Kuo, L.; Gardell, S. J.; Shafer, J. A.; Vacca, J. P. Metabolism-directed optimization of 3-aminopyrazinone acetamide thrombin inhibitors. Development of an orally bioavailable series containing P1 and P3 pyridines. *J. Med. Chem.* **2003**, *46*, 461–473.
- (29) Feng, D. M.; Gardell, S. J.; Lewis, S. D.; Bock, M. G.; Chen, Z. G.; Freidinger, R. M.; Naylor-Olsen, A. M.; Ramjit, H. G.; Woltmann, R.; Baskin, E. P.; Lynch, J. J.; Lucas, R.; Shafer, J. A.; Dancheck, K. B.

Chen, I. W.; Mao, S. S.; Krueger, J. A.; Hare, T. R.; Mulichak, A. M.; Vacca, J. P. Discovery of a novel, selective, and orally bioavailable class of thrombin inhibitors incorporating aminopyridyl moieties at the P1 position. *J. Med. Chem.* **1997**, *40*, 3726–3733.

(30) Lumma, W. C.; Witherup, K. M.; Tucker, T. J.; Brady, S. F.; Sisko, J. T.; Naylor-Olsen, A. M.; Lewis, S. D.; Lucas, B. J.; Vacca, J. P. Design of novel, potent, noncovalent inhibitors of thrombin with nonbasic P-1 substructures: Rapid structure–activity studies by solid-phase synthesis. *J. Med. Chem.* **1998**, *41*, 1011–1013.

(31) Sanderson, P. E. J.; Cutrona, K. J.; Dorsey, B. D.; Dyer, D. L.; McDonough, C. M.; Naylor-Olsen, A. M.; Chen, I. W.; Chen, Z. G.; Cook, J. J.; Gardell, S. J.; Krueger, J. A.; Lewis, S. D.; Lin, J. H.; Lucas, B. J.; Lyle, E. A.; Lynch, J. J.; Stranieri, M. T.; Vastag, K.; Shafer, J. A.; Vacca, J. P. L-374,087, an efficacious, orally bioavailable, pyridinone acetamide thrombin inhibitor. *Bioorg. Med. Chem. Lett.* **1998**, *8*, 817–822.

(32) Isaacs, R. C. A.; Solinsky, M. G.; Cutrona, K. J.; Newton, C. L.; Naylor-Olsen, A. M.; McMasters, D. R.; Krueger, J. A.; Lewis, S. D.; Lucas, B. J.; Kuo, L. C.; Yan, Y. W.; Lynch, J. J.; Lyle, E. A. Structure-based design of novel groups for use in the P1 position of thrombin inhibitor scaffolds. Part 2: N-acetamidoimidazoles. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 2062–2066.

(33) Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* **1999**, *285*, 1735–1747.

(34) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **2003**, *24*, 1999–2012.

(35) 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.

(36) 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.

(37) 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.

(38) Lill, M. A.; Danielson, M. L. Computer-aided drug design platform using PyMOL. *J. Comput.-Aided Mol. Des* **2011**, *25*, 13–19.

(39) 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.

(40) Sharp, K. A.; Nicholls, A.; Fine, R. F.; Honig, B. Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects. *Science* **1991**, *252*, 106–109.

(41) Sitkoff, D.; Sharp, K. A.; Honig, B. Correlating solvation free energies and surface tensions of hydrocarbon solutes. *Biophys. Chem.* **1994**, *51*, 397–403.

(42) Wang, R.; Wang, S. How does consensus scoring work for virtual library screening? An idealized computer experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422–1426.