Fragment-Based Computation of Binding Free Energies by Systematic Sampling

Matthew Clark,[†] Siavash Meshkat,* George T. Talbot, Paolo Carnevali,[‡] and Jeffrey S. Wiseman[†]

Locus Pharmaceuticals, 512 E. Township Line Road, Blue Bell, Pennsylvania 19422

Received April 8, 2009

A fragment-based method for computing protein−ligand binding free energies by systematic sampling has been developed. Systematic sampling of fragment−protein interactions in translational and rotational space is followed by de novo assembly of fragments into molecules and computation of binding free energies for the molecules with statistical mechanics. The rigorous sampling provides independence from the choice of initial binding pose and assembling fragments enables evaluation of binding of a large number of molecule poses with relatively little computation. The method allows a full sampling of possible conformations and avoids the "conformational focusing" problem associated with free energy methods that sample only limited conformational and orientation changes from a starting pose. The direct computation of the entropy loss upon assembling fragments into molecules is an innovation for fragment-based methods. The computed binding free energies are compared to calorimetric data for a series of ligands for the T4 lysozyme L99A mutant and binding constants for a series of p38 MAP kinase ligands. In both cases, the standard error of prediction is close to 1 kcal/mol.

INTRODUCTION

Increasing access to molecular diversity has been a major driving force for developing new technologies in medicinal chemistry over the last two decades, driven largely by the major advances in biology.[1] Diverse ligand structures are required, for example, in order to achieve activity across multiple protein families, presenting structurally diverse binding sites. Similarly, identifying sufficient diversity to identify novel, patentable chemical series for a target protein can be a significant challenge when high structural homology is broadly retained across a protein family, as is the case for the kinases, for example. Finally, molecular diversity has become a major factor in strategies to optimize in vivo efficacy while minimizing toxicity.[2] In response, major infrastructures for automated and combinatorial chemistry and ultrahigh throughput screening have been developed to increase the size of corporate screening libraries by an order of magnitude.[3] Given limitations on synthesizing, storing, and manipulating large physical compound collections, however, the focus of efforts to access additional diversity is shifting to methods that are more efficient than simply increasing the size of compound collections.

Fragment-based drug design is currently evolving as a major method for efficient access to diverse molecular structures.[4] The principle behind this method is that small ligands will be less discriminating between targets so that viable leads for new protein targets can be identified efficiently from small screening libraries.[5] There are limitations in the implementation of experimental fragment-based design, however, arising largely from the fact that small ligands typically bind more weakly to protein binding sites than fully optimized drug-like ligands. This requires new screening techniques tuned to identify weak ligands and generally utilize NMR or crystallography as the primary assay or as a supplementary method to verify that the binding modes for weak ligands are productive. The result is that significant physical infrastructures are required to implement these methods.[6] Current technologies are generally tuned to detect binding in the 0.1−10 mM range; ligands in the range 200−250 Da in size are required to achieve these potencies, approximately half the size of ligands in a classical screening library.

Computational methods would be highly compatible with fragment-based design; and augmenting the current experimental approach with the computational equivalent could, in theory, increase the efficacy of this approach while circumventing some of the limitations. First and foremost, there is no lower limit to binding potency for computational methods so that, in turn, very small fragments may be utilized. This would allow us to extend the principle of ubiquitous binding of smaller ligands to the extreme so that ligand binding efficiency[7] could be optimized at the fragment level for each local binding 'hot spot' in a binding site. Moreover, there would be no limitation on the physical properties, such as solubility, for a computational fragment, further increasing the effective diversity of the fragments themselves.

The most important requirement for an effective computational approach is high speed at low cost. Any method would have to match or exceed a screening rate of 10^3 fragments per day at a cost that is competitive with high-throughput crystallography and other assays tuned for fragment screening.[3] Meeting this requirement will introduce the second and most critical performance criterion of a computational method, namely the ability to reliably discriminate between active and inactive molecules in the virtual library that exceeds the discriminating power of common

* Corresponding author. E-mail: smeshkat@Locuspharma.com.

[†] Present address: Pharmatrope Ltd., 324 Croton Road, Wayne, PA 19087.

[‡] Present address: Complete Genomics Inc., 2071 Stierlin Court, Mountain View, CA 94043.

10.1021/ci900132r CCC: $40.75
Published on Web 07/17/2009

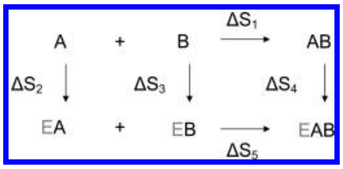

Figure 1. Thermodynamic cycle of fragments A and B binding to the protein E separately and in combination AB.

virtual screening approaches. In practice, the computations would assess binding of fragments at multiple binding hot spots in a protein binding pocket, and full-sized ligands would be designed by linking the best fragments together to form a full-sized ligand. For 1 000 fragments and 3−4 binding hot spots, this would be equivalent to assessing a virtual combinatorial library of 10^9-10^{12} molecules. These numbers represent the potential for access to a very high diversity by this approach. If the second performance criterion is met, then only a very small fraction of such a large virtual library needs to be synthesized.

The purpose of this paper is to introduce a core method for computing fragment binding affinities that meets the requirements of high speed and low cost and has the promise of reliably distinguishing active from inactive compounds in very large virtual libraries. A fully implemented computational method must address a number of computationally intensive problems: calculation of the gas-phase binding free energy for fragments, correction of the gas-phase affinity to aqueous solution, calculation of the binding free energy for a molecule from the binding free energy of its component fragments, and allowance for conformational flexibility in both the ligand and protein. The core method provides calculation of binding free energies for fragments and molecules subsequently assembled from the fragments. We show, in addition, that a fragment-based approach is inherently efficient for dealing with conformational flexibility in the assembled molecules. Although a rudimentary treatment of solvation energy and protein flexibility are used in the core method, the method is fast enough that more sophisticated treatments can be considered for the future, which is expected to further improve the results. A historical perspective on the calculation of fragment and molecule free energies is provided in the following.

The distinguishing feature of a fragment-based computational method is the estimation of binding free energies for whole molecules from the binding free energies of their component fragments. Not all methods that might be used to compute fragment binding affinities are compatible with subsequent efficient calculation of the binding affinity of the whole molecule, so this aspect of the calculation must be considered first. The fact that binding energies are not additive has long been appreciated. The difference between the total binding energy and the sum of the components was termed the “connection Gibbs energy” by Jencks.[8,9] A major source of this connection energy is the entropy change that occurs when two fragments are linked to form a molecule. The largest component of this change is the loss of translational and rotational freedom associated with combining two bodies into one, but the molecule also loses conformational freedom when it binds to a protein, and this entropy loss is not easily related to binding of the individual fragments.[9,10] The entropy changes involved are illustrated in Figure 1. The entropy change when connecting bound fragments, $\Delta S_{5,}$ is particularly difficult to estimate. Methods to calculate these entropy changes from purely experimental data have not yet succeeded, and empirical approximations are generally used instead.[10] Since the ultimate goal is to compute the entropy component, ΔS_4, of the binding free energy of the whole molecule and the entropies ΔS_2 and ΔS_3 will be inherently available from the fragment binding free energies, the additional requirement for the fragment free energy calculations is that they must be capable of yielding the difference between the entropies of the whole molecule in solution and bound to the protein, ΔS_1 and ΔS_5.

Several possible approaches can be considered to compute fragment−protein binding affinities. Approximate methods to explore ligand−protein binding, such as docking, have shown great utility in quickly screening large libraries of molecules and scoring their ability to fit in a given binding site. While these methods can reproduce binding poses accurately, they have performed poorly at predicting binding affinities.[11] To be predictive, methods must focus on computing the binding free energy, which is directly related to the physically measurable K_d; significant effort has been devoted over the last decades to developing fast, robust methods to meet this goal.[12−15] Full binding free energy calculations can be classified into two general approaches. Pathway approaches simulate the transition from the bound to the unbound state by either alchemical changes or actual motion of the molecule between the states. The potential of mean force method (PMF) is one pathway approach, which computes binding by integrating the work required to pull the bound ligand out of the binding site into bulk solvent. Free energy perturbation (FEP) methods are a second example of the pathway class. These methods alchemically grow or remove ligands in the binding site and have been among the most used over the past decades.[16] FEP methods include the “double annihilation”[17] and “double decoupling”[18−20] methods. The FEP and other thermodynamic integration methods include no approximations other than those inherent in the energy function and the choice of sample size.

The second family of free energy methods, called “end point” methods, computes only properties of the bound system and of the free ligand and derives the free energy of binding from the difference. The simplest of these is the linear interaction energy (LIE) model that combines QSAR principles with molecular mechanics interaction energies to create a predictive relationship that relates the average van der Waals and electrostatic energies to the observed binding energies. The molecular mechanics/generalized Born-surface area models (MM/GB-SA) or Poisson−Boltzmann surface-area models (MM/PBSA) are more rigorous end point methods. These methods use molecular dynamics to sample the binding poses available in the protein while computing the solvation free energy using the GB-SA or PB-SA methods.[21] A third example of this class is the grand canonical free energy simulation. This method computes the free energy of fragments binding to proteins by equilibrating concentrations between a reference state and a simulation cell that includes fragments bound to a protein.[22,23]

Each of these free energy methods fulfills the first criterion for our computational method by predicting binding affinities for a variety of ligands and proteins. The molecular mechanics model provides binding free energies with reported errors in the vicinity of 1 kcal/mol, for example.[9,24,25] With the exception of the grand canonical method, however, these methods all begin with a starting pose defining the protein−ligand binding mode and compute the free energy of that pose. Although this approach greatly simplifies the complexity of the computational task, it can be subject to error if multiple low-energy poses are accessible, and a recent study estimated that this error can be on the order of 5 kcal/mol.[26−28] More importantly for fragment-based methods, limiting the system to a single pose does not provide the information needed to compute the entropic changes associated with fragment assembly.

The grand canonical method, in contrast, was designed specifically to efficiently sample multiple poses and avoid any errors that might arise from assuming a single starting pose for the ligand.[22,23] The method was also designed to be fragment-based and provides the necessary data to compute the entropy difference between individual fragments bound to a protein and the entropy of the assembled molecule in the same binding site. Because the method efficiently samples all available poses in the binding site at very high resolution, this entropy difference can be easily computed simply by comparing the poses available to the fragments individually vs as a moiety of the whole molecule. The grand canonical method does not, however, allow calculation of the complete Gibbs connection energy. In order to accomplish this, it is also necessary to compute the entropy change associated with combining fragments into molecules in the absence of protein (ΔS_1 in Figure 1), and the grand canonical method is not well suited for this purpose.

Therefore, we have created a new method for computing protein−ligand binding free energies that fully meets the requirements of computing both individual fragment binding affinities and approximating the Gibbs connection energy. This method does not require prior identification of the binding mode, can efficiently scan a selected binding region for tight interactions, and computes binding free energies for a very large number of molecules in a large number of poses and conformations with modest computational effort. The method first performs complete systematic sampling of chemically significant fragments in the binding site. The fragments are then assembled into molecules, and the ligand−protein binding free energy is integrated using the set of assembled poses bound to the protein. Finally, the integration is repeated for the fragment poses assembled in the unbound reference state. Comparison of the results for the bound and unbound states provides a good approximation of the Gibbs connection energy. The integration process implicitly includes the loss of degrees of freedom created by binding the fragments together and, thus, provides the correct binding entropy for the assembled molecules. Rigid fragment conformations are used in the free energy simulation. If a fragment itself is conformationally flexible, then it is modeled as a collection of multiple rigid conformers.

Conformational flexibility at the bonds formed between fragments is inherently accounted for in the integration over fragment poses. A high resolution is required in order to ensure visiting all of the important poses, resulting in the need for sophisticated handling of large amounts of data.

The goal of this study is to describe the implementation of this new approach for fragment-based drug design. In order to demonstrate its efficacy in predicting the ligand binding affinities, we compare computed to observed affinities for ligands of bacteriophage T4 lysozyme, which has played a longstanding role as a test bed for stability and structure studies.[29−32] The T4 lysozyme L99A system is a nearly ideal test system for computing ligand binding due to the small size of the artificial binding pocket and availability of a series of crystal structures along with calorimetric binding data. The system has been used for examining both free energy computations and docking so that comparisons with other methods are possible.[33−36] The T4 L99A binding pocket is completely enclosed by the protein and admits only small ligands. These ligands can be assembled with one to three fragments and have limited flexibility. The protein has been shown to be reasonably rigid, and no water molecules are observed in the binding site with ligands. The generally hydrophobic ligands and binding pocket do not require complex solvation treatments, further simplifying the study. There is a wealth of calorimetry and crystallographic data determined for T4 ligands that can be compared to the computed results. In addition, data for T4 ligands have been widely studied by other groups and, therefore, allow comparison of a variety of free energy computation methods.

In addition, we validate the method using experimental data for p38 mitogen-activated protein kinase, which is a regulator of the signal transduction cascade leading to release of tumor necrosis factor-α and interleukin-1β in inflammatory diseases.[37] A series of ligands, based on a pyrazole-urea scaffold, substituted at three points was recently reported for the allosteric binding site of p38.[45] These ligands form a congeneric series with sufficient diversity, therapeutic relevance, and range of activities to be a test of the method that is interpretable in terms of drug discovery.

METHODS

Overview. The free energy calculation is carried out in three steps. The first is to perform systematic sampling in the six translational and rotational dimensions for rigid fragments, computing the molecular mechanics force field energy between the fragment and a protein for each pose. The sampling is carried out in a volume selected to encompass a protein binding site. A set of rotations is selected to provide complete rotational sampling with a minimum number of poses. For fragments with internal conformational flexibility, for example cyclohexane, the relevant conformations are simulated separately as rigid conformers, and the results for the conformers are then pooled to compute the free energy for the fragment. The relative energies of each conformation are computed for use in eq 8a, as described below. The computation of the free energies and analysis of binding modes of fragments requires approximately 30 min per fragment on a 2.6 GHz Intel Core2 Duo 6700 processor at 0.2 Å translational resolution, generating approximately 100 MB of data per fragment.

The next step is assembly of the fragments into more complex molecules. Ethylbenzene, for example, may be constructed from ethane and benzene fragments. Poses of the two fragments are first sampled independently. After sampling the individual fragments, the geometries of their poses are analyzed to find all pairs of benzene and ethane poses that place a carbon of ethane within a predetermined

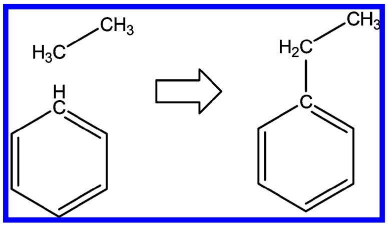

Figure 2. Example of constructing ethylbenzene from a pose of ethane and benzene. A hydrogen atom is removed from benzene, and a terminal hydrogen from ethane and the two carbon atoms are bonded.

distance and angle tolerance to form a bond to a carbon of benzene (Figure 2).

The final step is computation of the binding free energies of the assembled molecules using the poses that were assembled and a comparable set of assembled poses that form the unbound reference state.

Translational Sampling. The sampling of fragment translations is straightforward. We successively translate the centroid of the fragment to points of a three-dimensional rectangular grid. The three translational resolutions are kept equal, so that the translational sampling grid consists of cubes with sides of length Δ_T. Sampling resolution is characterized by the maximum and the average distances of any point within a grid cube and the closest grid point, which are at the corners of the cube. For any point in space, the distance to the closest grid point can never be larger than half diagonal of the cube $\sqrt{3}\Delta_T/2$.

Rotational Sampling. Complete sampling is achieved by combining each of the translation vectors with a set of fragment rotations. The process of selecting rotations starts with the generation of a large superset, S_0, consisting of N_R fragment rotations selected randomly with a uniform distribution in quaternion space.[38] We then select a subset, S_1, of the N_R fragment rotations to be used in the sampling from S_0 in such a way that the size of S_1 is minimized while at the same time the distance between any two rotations falls within the desired sampling resolution. The algorithm is described in the Supporting Information. We define the distance between two fragment rotations as the atomic root-mean-square displacement of atoms generated when the fragment is moved from the first rotation to the second one. With this metric, the distance between two rotations takes into account the shape of nonspherical fragments and does not depend solely on the angle between the two rotations.

Computation of Interaction Energy by Grid Interpolation. Grid interpolation provides a fast computation of the interaction energy between the fragment and the protein for each pose with little loss of accuracy. In this study we use nonbonded terms of the AMBER force field.[39] The protein charges are taken from the published AMBER99 force field, and fragment charges are computed using the CHELPG method.[40]

Without making any approximation, the interaction energy can be rewritten as eq 1, where $\varphi(\mathbf{r})$ and $\Psi_a(\mathbf{r})$ are "potential" scalar fields, independent of the positions of the fragment atoms (eqs 2 and 3):

$$E = \sum_a [q_a\phi(\mathbf{r}_a) + \psi_a(\mathbf{r}_a)] \tag{1}$$

$$\phi(\mathbf{r}) = \sum_b \frac{kq_b}{(\mathbf{r} - \mathbf{r}_b)^2} \tag{2}$$

$$\psi_a(\mathbf{r}) = \sum_b \varepsilon_{ab}\left[\frac{\sigma_{ab}^{12}}{(\mathbf{r} - \mathbf{r}_b)^{12}} - 2\frac{\sigma_{ab}^{6}}{(\mathbf{r} - \mathbf{r}_b)^{6}}\right] \tag{3}$$

The number of distinct $\Psi_a(\mathbf{r})$ fields is equal to the number of distinct AMBER atom types in the fragment. The above expressions for the interaction energy can be evaluated very rapidly, if the required values of $\varphi(\mathbf{r})$ and $\Psi_a(\mathbf{r})$ at the positions of the fragment atoms are available. We compute values of $\varphi(\mathbf{r})$ and $\Psi_a(\mathbf{r})$ on a three-dimensional rectangular grid with resolution Δ_F. This grid is similar but distinct from the grid used to sample translations and rotations described in the previous section. In particular, the resolutions for the sampling and energy grids are not required to be equal. In this study, the energy grid was 0.15 Å. Values of $\varphi(\mathbf{r})$ and $\Psi_a(\mathbf{r})$ at atomic positions are then computed by trilinear interpolation of the values at the eight corners of the grid cell containing each fragment atom. The energy evaluation time increases proportionally to the number of atoms in the fragment but is insensitive to the number of atoms in the protein. While we used a rigid protein in the study, the grid method can be generalized by recomputing the energy grids as the protein moves. The method can also be generalized so that the free energy of binding can be computed over a representative set of protein conformations.

Computation of Fragment Free Energy. Fragment sampling produces a large list of poses and their corresponding energies E_i. Since the procedure used to construct translations and rotations provides an essentially uniform coverage of the fragment configuration space, it is a reasonable approximation to replace configuration integrals that appear in statistical mechanics equations with sums over the computed poses. For example, we can compute the partition sums Z for the bound state and Z_0 for the unbound state and the Helmholtz binding free energy by a simple sum over poses using eqs 4–6. Since changes in pressure/volume during binding are negligible, the Gibbs free energy of binding, ΔG, is approximately equal to the Helmholtz free energy in eq 6.

$$Z = \sum_i e^{-E_i/kT} \tag{4}$$

$$Z_0 = n_R V_0/\Delta_T^3 \tag{5}$$

$$\Delta G = -kT \ln(Z/Z_0) \tag{6}$$

An energy cutoff is selected, such that high-energy fragments with negligible contribution to the sum in eq 4 are discarded from the samples. A cutoff of 0 kcal/mol was used in this study. This reduced the number of poses to be considered by 3−6 orders of magnitude, without appreciably changing the partition sum.

In the reference, an unbound state of 1 mol/L concentration, all poses have zero interaction energy with a volume per pose of V_0 = 1 660 Å^3. The number of poses in the reference state is, thus, $n_R V_0/\Delta_T^3$, for a box of size $\Delta_T \times \Delta_T \times \Delta_T$. This gives the partition function, Z_0, for the reference state in eq 5. The binding enthalpy is computed with eq 7, and the entropy can be computed from the difference between free energy and enthalpy.

$$\Delta H = \frac{1}{Z}\sum_i E_i e^{-E_i/kT} \tag{7}$$

Computation of Free Energy of Molecule. The binding free energy of the molecule is computed as the difference between its energies in the bound and unbound states, as described below.

Computation of Energy of Molecule in the Bound State. The next step is assembly of the fragments into more complex molecules. For example, ethylbenzene, constructed from ethane and benzene fragments, may be simulated independently as shown in Figure 2. After sampling the individual fragments, the geometries of their poses are analyzed to find all combinations where a carbon of ethane is at the appropriate distance and angle for bond formation to a carbon of the benzene, within a selected distance and angle tolerance of the ideal values. The tolerances used in this study are 0.5 Å and 15°, chosen to be approximately 2-fold larger than that of the sampling resolution. As shown in eq 8a, fragment energy E_f is the sum of the fragment−protein interaction energy, $E^{\text{fragment−protein}}$, the fragment solvation energy, $E^{\text{solvation}}$, the fragment internal conformational strain, $E^{\text{fragment strain}}$. The total energy of each assembled molecule pose is computed by adding a sum over fragments f and a sum over fragment pairs f,g, as shown in eq 8b. The strain introduced in the fragment assembly of a fragment pair is denoted as $E^{\text{interfragment strain}}$.

$$E_f = E_f^{\text{fragment−protein}} + E_f^{\text{solvation}} + E_f^{\text{fragment strain}} \tag{8a}$$

$$E = \sum_f E_f + \sum_{f,g} E_{f,g}^{\text{interfragment strain}} \tag{8b}$$

The fragment strain energy is used when linking to a set of conformations of structurally identical fragments, such as the boat and chair forms of cyclohexane. It contains the inherent strain in the conformation of the fragment used to construct a molecule. The interfragment strain energy is computed using the force field terms corresponding to the new connection made between the fragments − the bond stretch, angle bend, torsion, and Coulomb and van der Waals nonbonded terms. In these terms, the bond stretch and angle bend energy contributions are attenuated to compensate for nonideal bond lengths and angles created by grid sampling. The torsion and nonbonded energies are not scaled. We empirically chose a linear scaling of bond and angle terms with weights of 0.05 and 0.15, respectively.

The binding free energy of a fragment is computed using the Boltzmann sum of the poses in the bound state, normalized by the sum of poses for the reference state as in eq 6. The computation of Z for the bound states is straightforward since the presence of a protein sterically limits the poses that can be assembled in the binding site. However, in the unbound case required for Z_0, this reduction is not present, therefore, the number of assembled poses is too large for direct computation.

Computation of the Reference State of Two Joined Fragments. For any set of fragments U in the reference state, we let $\nu(U)$ denote the number of poses and $Z_0(U)$ denote the partition sum. For two neighboring fragments A and B, we let $\nu(\text{B|A})$ and $Z_0(\text{B|A})$ represent the number of poses and partition sum of B, given a fixed pose of fragment A.

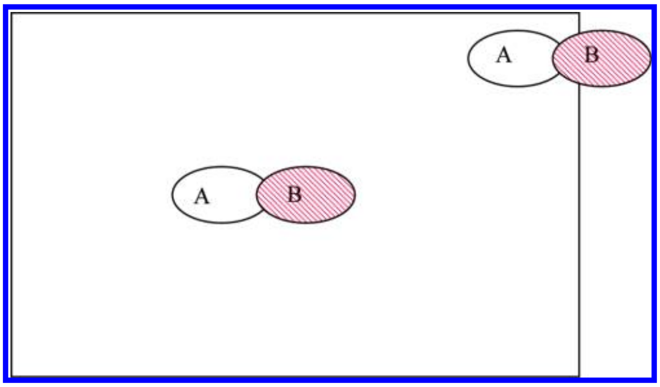

Figure 3. The molecule pose on the upper right places fragment B centroid outside the box, therefore, it is not included in the reference state calculation. Boundary correction factor μ for the two fragment join of A and B inside the sampling box accounts for poses that cannot fit in the sampling box.

To compute the unbound state of a two-fragment molecule AB, a fragment A is joined to a neighboring fragment B. We fix the fragment with the largest number of rotational poses (say A) at the center of the sampling box. To compute $\nu(\text{B|A})$, we enumerate the poses of the other fragment (B) that can connect to the fixed fragment by forming a bond, within the given bond distance and angle criteria, with strain energy below 50 kcal/mol. The computed poses of fragment B that can mate with the fixed pose of fragment A are stored for downstream computations.

Given the above, we then compute the number and partition sum of the unbound states of submolecule AB using eqs 9 and 10. The multiplicative factor μ is explained below.

$$\nu(\text{AB}) = \mu\nu(\text{A})\nu(\text{B|A}) \tag{9}$$

$$Z_0(\text{AB}) = \mu\nu(\text{A})Z_0(\text{B|A}) \tag{10}$$

The computation of $Z_0(\text{B|A})$, the partition sum of all poses of B given a fixed pose of A, is performed with eq 11:

$$Z_0(\text{B|A}) = \sum_{\text{B}} e^{-E_{\text{AB}}/kT} \tag{11}$$

The energy of each pose pair, E_{AB}, in eq 11 is computed in terms of the fragments and molecule conformational strain energies and the number of bonds connecting to fragment A (n_A) and fragment B (n_B) by eq 12:

$$E_{\text{AB}} = \frac{1}{n_{\text{A}}}E_{\text{A}}^{\text{fragment strain}} + \frac{1}{n_{\text{B}}}E_{\text{B}}^{\text{fragment strain}} + E_{\text{AB}}^{\text{interfragment strain}} \tag{12}$$

The multiplicative factors $1/n_{\text{A}}$ and $1/n_{\text{B}}$ are used to avoid multiple counting of the internal fragment strains, when two fragment reference states are combined. For example, fragment A strain appears in n_{A} energy terms of the form E_{AB}, which are added as a result of multiplying exponentials.

The factor μ in eq 10 corrects the boundary effect, where, for some poses of A, the centroid of the connected fragment B would be outside the sampling box and, therefore, should not be included in the reference state. This is illustrated in Figure 3. The factor μ is estimated as 1 − fragment radius/Δ_T.

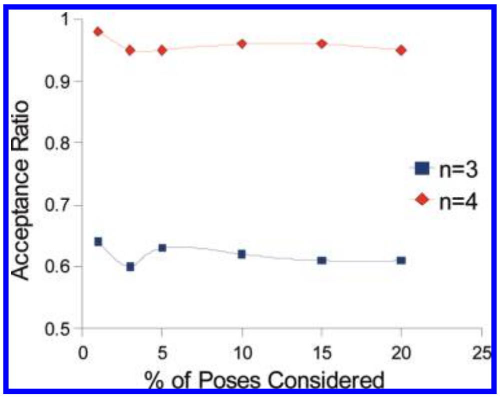

Figure 4. The van der Waals clash acceptance ratio for 3 and 4 fragment joins.

Computation of Reference State of More than Two Joined Fragments. In the previous sections, we considered the interaction of two fragments immediately adjacent to each other. In this section, we consider more remote interactions that arise from adding additional fragments. For connecting multiple fragments, let A and B represent two sets of fragments that share a common fragment C. In this case, the pose of fragment C at the center of the box is the starting point. All poses of fragments A and B that can connect to the fixed fragment C, meet the geometric requirements for bonding and do not have van der Waals clashes that are enumerated. The reference state integral Z_0 is approximated by estimating the fraction of poses that have steric overlaps via a van der Waals clash acceptance ratio, ρ(A,B), the fraction of poses that do not clash.

The total number of poses and partition sum for the submolecule AB is computed using eq 13 and 14, respectively:

$$\nu(\mathrm{AB}) = \rho(\mathrm{A,B})\frac{\nu(\mathrm{A})\nu(\mathrm{B})}{\nu(\mathrm{C})} \qquad (13)$$

$$Z_0(\mathrm{AB}) \approx \rho(\mathrm{A,B})\frac{Z_0(\mathrm{A})Z_0(\mathrm{B})}{\nu(\mathrm{C})} \qquad (14)$$

The full enumeration of possible unbound poses results in a very large number. Even for four fragments, this number can well exceed 10^{20} poses. For this reason, a random subset of poses is sampled. Figure 2 shows the acceptance ratio ρ, as a function of the percent of poses considered of each fragment.

The acceptance ratio converges quickly after 5% of poses of each fragment are considered, demonstrating that the random sampling of a subset of unbound poses provides a stable and accurate estimation of the reference state. Due to the combinatorial nature of this step, using a sampling of 5% of reference poses can reduce computational time considerably with a small effect on the free energy. The approximations were tested by sampling methane in an 8 Å cubic box without protein and assembling ethane from two methane molecules. The bond distance and angle tolerances were 0.5 Å and 15°, respectively. Since there is no fragment−protein interaction energy in this experiment, the number of poses in the reference state should equal the assembled poses. The method assembled 1 183 047 172 poses of ethane and estimated a reference state of 1 303 240 000 poses, resulting in an error of 10%. Since the free energy depends on the log of this number, the corresponding error in free energy is about −0.02 kcal/mol.

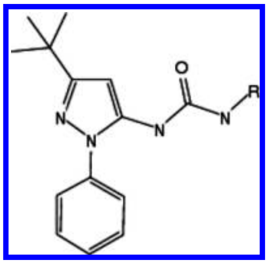

Figure 5. Scaffold for p38 ligands.

T4 Lysozyme and Fragment Preparation. A series of cocrystal structures and thermodynamic binding data are available for T4 lysozyme L99A, with an artificial binding pocket created via the mutation of leucine 99 to alanine.[41,42] The crystal structure cocrystallized with *n*-butyl benzene, pdb reference 186L, has the largest binding pocket and was used for this study. The ligands were prepared, as previously described,[21] by energy minimization with MacroModel,[43] computation of charges by the CHELPG method using Gaussian 98,[44] and then computation of the solvation free energy using the GB/SA method implemented in MacroModel. The sampling was performed at a Cartesian resolution of 0.2 Å and a minimum rotational resolution of 0.333 Å rms, which yielded an average rotational resolution of 0.14 Å rms after selection of rotations. The sampling region was a cube 8 Å on a side, centered on the T4 binding cavity. Poses with energies higher than −5 kcal/mol were discarded to help reduce the total number of poses to process. The gas to aqueous solvation energy was added to the final energy to correct for the solvation energy of the ligands, as described previously.[6]

p38 MAP Kinase and Fragment Preparation. A series of compounds has been reported that bind in the allosteric site of p38 MAP kinase.[45] The protein was prepared by adding hydrogen atoms and terminal groups to the 1KV2 structure.[46] The activation loop from another published p38 structure, 1A9U,[47] was added to the model and subjected to 1 ns of molecular dynamics in torsion space, while the rest of the protein was restrained from motion. In addition, water molecules conserved across several protein structures were included in the model with the hydrogen positions optimized with energy minimization. The fragments necessary to construct the scaffold, 3-*tert*-butyl-1-*H* pyrazole, urea, and benzene as well as the fragments necessary to construct a series of ligands substituted at the urea were prepared, as described above. The scaffold of these compounds is shown in Figure 5. The list of fragments used to substitute at the urea is listed in Table 3. The sampling region was a parallelepiped centered on the crystallographic position of the urea nitrogen, measuring 15, 15, and 12 Å in the *x*, *y*, and *z* dimensions. The remaining sampling parameters were the same as described for T4 lysozyme.

RESULTS

Sampling Efficiency. If the volume of a binding mode, measured in the six dimensions of translation and rotation is small relative to the sampling resolution used, then the systematic sampling procedure may not sample it. The small binding pocket in T4 lysozyme L99A was chosen as a

Table 1. Computed and Experimental Energies in kcal/mol for T4 Lysozyme Ligands[a]

			calcd ΔG			
molecule	calcd ΔH	calcd $T\Delta S$	uncorrected	solvation corrected	solvation and conformational energy corrected	obsvd ΔG
			One Fragment			
o-xylene	−18.9	−10.4	−8.6	−9.9	−9.9	−4.6
p-xylene	−17.6	−10.7	−7.0	−8.5	−8.5	−4.7
m-xylene	−19.6	−11.0	−8.5	−10.1	−10.1	−4.8
indole	−22.2	−10.4	−11.8	−8.7	−8.7	−4.9
indene	−21.0	−10.7	−10.2	−10.7	−10.7	−5.1
benzene	−16.1	−7.2	−8.9	−9.4	−9.4	−5.2
benzofuran	−22.3	−10.1	−12.2	−11.0	−11.0	−5.5
toluene	−19.1	−9.1	−10.1	−11.1	−11.1	−5.5
thianaphthene	−21.3	−10.5	−10.8	−10.8	−10.8	−5.7
			Two Fragments			
2-ethyl toluene	−20.0	−13.8	−6.3	−8.9	−11.0	−4.6
3-ethyl toluene	−18.7	−14.2	−4.6	−7.2	−9.4	−5.1
4-ethyl toluene	−22.8	−14.3	−8.5	−11.2	−11.8	−5.4
ethyl benzene	−22.6	−11.8	−10.8	−12.9	−15.1	−5.8
n-propyl benzene	−25.1	−14.1	−11.0	−13.2	−14.9	−6.6
			Three Fragments			
iso-butyl benzene	−25.7	−19.6	−6.0	−10.4	−15.1	−6.5
n-butyl benzene	−26.1	−19.0	−7.1	−11.0	−15.0	−6.7

[a] Table is separated by the number of fragments used to construct each molecule.

stringent test of the systematic sampling method to achieve the required sampling efficiency, which we describe using three metrics.

First, the method used to sample rotations is designed to increase sampling efficiency by emphasizing the most important rotational poses based on the shape of the fragment. For example, indole sampled on a simple grid with 15° increments of Euler angles provides 6 912 rotations and results in a mean rms displacement of 0.32 Å, whereas a comparable number of rotational poses, 7 274, with the Monte Carlo algorithm used in this study gave a minimum rms displacement of 0.5 Å and a mean rms of 0.22 Å. Thus, the algorithmically selected poses produce a mean rms difference one-third better than those of the systematic sampling of Euler angles with nearly the same number of rotations. Although this difference is small, the number of rotational samples required for a given resolution is a factor in all six dimensions of translational and rotational space and increases as the sixth power of the resolution. The observed difference in mean resolution, therefore, represents an approximately 10-fold improvement in performance. This advantage over the Euler angles is even higher for molecules that have higher symmetry or are more nonspherical. For example, the proposed method would omit sampling rotations around the long axis of acetylene, since there is no rms movement of atoms for those rotations.

The second metric is the probability of finding a single pose, for example, the lowest enthalpy pose, at a given resolution. Since the probability of sampling a given pose *exactly* is infinitesimally small, we will use the thermal de Broglie wavelength to represent the uncertainty in the position of a single pose and the maximum meaningful resolution. The wavelength for benzene is 0.1 Å at 300 K, for example, so that we are approaching the maximum theoretical resolution for small fragments at the 0.2 Å translational and 0.14 Å average rotational resolutions used in these experiments. Since the number of poses sampled and the computation time increases as does the cube of the box size, we are also approaching the practical limit for the resolution.

Finally, sampling resolution can be cast in terms of the entropy of the binding mode; there is a critical entropy below which the binding mode cannot be detected. The lowest possible entropy for a single fragment is that of a single pose in the binding site. On the other hand, the reference state consists of $n_R V_0 / \Delta_T^3$ poses (eq 5). The entropy difference between the single pose and the reference state is given by eq 15:

$$T\Delta S_{\text{critical}} = kT \ln(\Delta_T^3/n_R V_0) \qquad (15)$$

The systematic sampling procedure will not be able to detect modes with entropy lower than this critical value. For typical values $\Delta_T = 0.2$ Å and $n_R = 10^4$, the critical value is $T\Delta S = -12.8$ kcal/mol at 300 K. Data presented below shows that the computed entropy reaches or exceeds this limiting entropy for the majority of ligands, particularly when conformational rotations must be accounted for. In principle in such cases, simply increasing the sampling would lower the limiting entropy to the necessary levels; for the T4 lysozyme example, this would require more than a 1 000-fold increase in sampling, which would increase computational times from CPU hours to CPU months. As discussed further below, it is much more efficient in these cases to assemble molecules from fragments, each of which individually does not exceed the entropy limit.

Simulation of Binding to T4 Lysozyme. Figure 6 shows the correlation of computed and observed binding free energies for small, rigid ligands, which were simulated as single fragments binding to the crystal structure 186L, the cocrystal structure of T4 lysozyme with *n*-butyl benzene. The computed free energies are lower than those of the observed binding energies; however, they are linearly correlated with a slope of 0.34 and an intercept of 4.2 kcal/mol. This offset is similar in magnitude to the strain energy of the valine

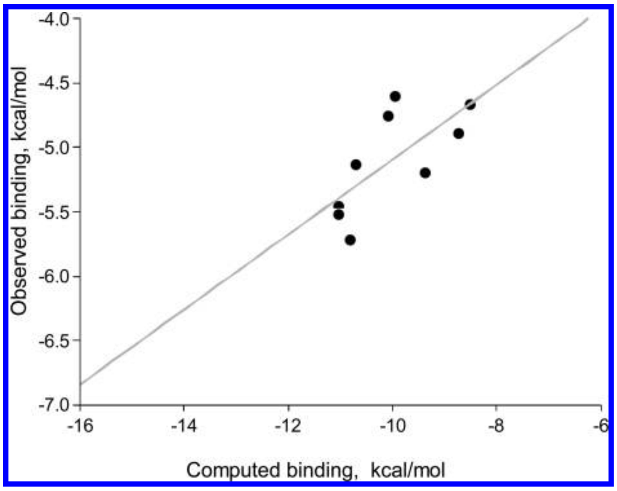

Figure 6. Computed vs experimental binding free energies for rigid compounds benzene, benzofuran, indole, indene, toluene, thianaphthene, *o*-xylene, *m*-xylene, and *p*-xylene, with T4 lysozyme.

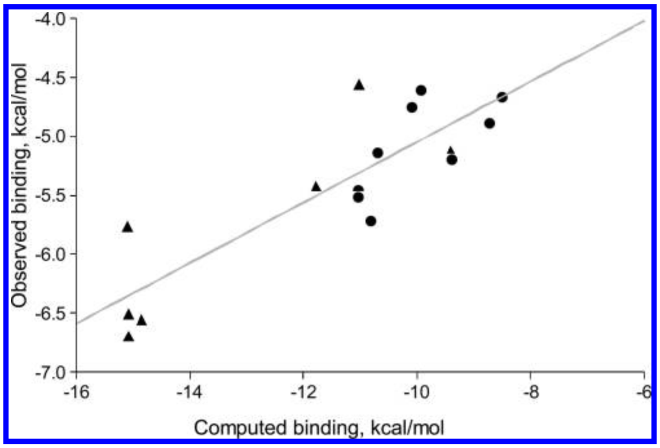

Figure 7. Computed vs experimental binding free energies for rigid and flexible compounds with T4 lysozyme. Circles represent single fragment compounds, triangles represent compounds assembled from multiple fragments.

rotamer flip required to form the pocket.[28,48] The r-squared of the correlation for this set of nine compounds is 0.49, and the standard error of prediction is 0.99 kcal/mol. The error of prediction is close to the lower limit of reliability of the molecular mechanics method.[24,25]

The rest of the compounds were assembled using multiple fragments with geometric tolerances of 0.5 Å bond tolerance from a 1.54 Å carbon−carbon bond length and an angular tolerance of 15° from the ideal bond angle. The compounds *n*-butyl benzene and *iso*-butyl benzene were constructed from three fragments {benzene, ethane, ethane} and {benzene, ethane, methane}, respectively. The ethyltoluenes were constructed from toluene and ethane, and ethylbenzene was constructed from benzene and ethane. Figure 7 shows the computed vs experimental results for ligands assembled from multiple fragments superimposed on the correlation for single-fragment ligands. The r-squared is 0.74, and the error of prediction is 1.22 kcal/mol. The fact that the data for single- and multiple-fragment ligands lie on the same correlation line is a validation of the fragment assembly algorithm. Details for the individual compounds are shown in Table 1.

Interfragment energy terms were required for good results. Figure 8 shows the energy results without the incorporation of the interfragment conformational energy. The two outliers when interfragment energies are not used are *n*-butyl benzene and *iso*-butyl benzene, which are constructed from three fragments each. In addition to the known ligands, a series of decoy compounds known to bind more weakly to T4 lysozyme were simulated (Table 2).[6,49] The affinity for quinoline and naphthalene is computed to be at the lower range of the observed binding. The binding affinity for naphthalene has been measured by melting point, and naphthalene is much less binding than the xylenes.[42] The other smaller molecules are computed to have much lower affinity than those observed in Table 1.

Table 2. Computed Energies in kcal/mol for Non-Binding Decoy Ligands to T4 Lysozyme

molecule	calcd ΔH	calcd TΔS	calcd ΔG, uncorrected	calcd ΔG, solvation corrected
ethanol	−12.9	−7.2	−5.7	−1.7
furan	−13.0	−6.1	−7.0	−5.9
methanol	−9.3	−5.0	−4.3	0.6
naphthalene	−18.8	−11.3	−7.6	−8.1
pyridine	−16.5	−7.2	−9.3	−7.1
quinoline	−21.9	−11.1	−10.8	−8.9

p38 MAP Kinase Ligand Binding. A series of p38 ligands[45] was assembled from four fragments as described (Figure 5). Approximately 2 × 10^6 poses were stored for each of the 66 fragments simulated. This resulted in a combined data size of 5.8GB. In order to reduce the computational time for fragment assembly, the small urea fragment was constrained to be within a 2 Å radius of its crystallographic position near Glu71. With this constraint, the average time to assemble each of the 44 molecules from the fragment data was 12 min.

Comparison of the computed and observed binding free energies for the p38 ligands is shown in Figure 9. As in the T4 study, the computed binding free energies are much lower than the observed values; however, there is a strong correlation between them. The r-squared for the relationship is 0.63, with a 0.88 kcal/mol standard error of prediction. The details for each compound derived from the scaffold shown in Figure 5 are shown in Table 3. As observed for

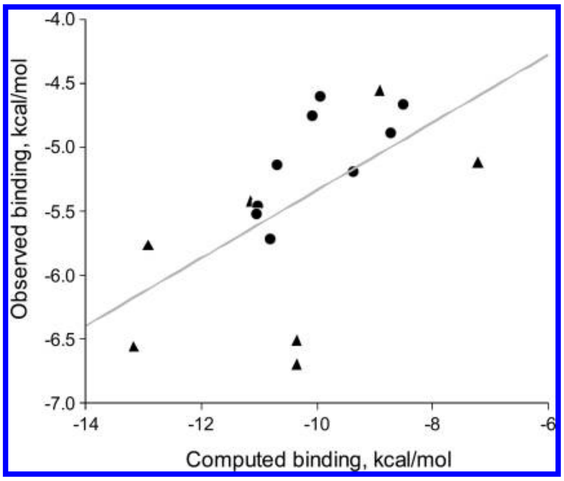

Figure 8. Computed (omitting interfragment conformational energy correction) vs experimental binding free energies for rigid and flexible compounds with T4 lysozyme. Circles represent single fragment compounds, and triangles represent compounds assembled from multiple fragments. These data demonstrate the improvement obtained from inclusion of interfragment conformational energy.

Table 3. Computed and Experimental Energies in kcal/mol for a Series of Ligands Reported For p38 MAP Kinase;[45] Parent Structure Shown in Figure 5

				calcd ΔG			
compound[a]	R	calcd ΔH	calcd $T\Delta S$	uncorrected	solvation corrected	solvation and conformational energy corrected	observed ΔG[b]
16	4-chlorobenzene	−94.2	−20.2	−74.0	−74.0	−42.6	−11.0
40	2-indane	−101.2	−22.4	−78.9	−78.9	−38.4	−10.7
46	benzene	−89.0	−18.9	−70.1	−70.1	−39.4	−10.7
65	H	−72.1	−15.6	−56.6	−56.6	−22.3	−7.0
66	cyclohexane	−91.2	−21.6	−69.6	−68.7	−34.5	−9.4
67	2-pyridine	−89.9	−19.5	−70.3	−70.3	−34.9	−8.7
69	3-pyridine	−90.5	−18.8	−71.7	−71.7	−37.8	−8.0
70	3-aniline	−91.2	−20.8	−70.5	−70.5	−39.3	−9.5
71	4-aniline	−94.5	−21.3	−73.3	−73.3	−41.2	−8.8
72	3-(1,2 dimethyl benzene)	−95.9	−20.3	−75.6	−75.6	−44.5	−11.6
73	toluene	−95.1	−21.2	−73.9	−73.9	−41.3	−10.0
74	ethylbenzene	−80.5	−22.2	−58.3	−58.3	−28.0	−9.0
75	1-naphthalene	−102.9	−20.8	−82.1	−82.1	−48.4	−11.3
76	2-naphthalene	−93.8	−21.7	−72.1	−72.1	−40.9	−11.3
77	1-indane	−106.1	−22.0	−84.2	−84.2	−43.8	−11.0

[a] Compound numbering from Regan et al.[45] [b] Observed ΔG computed from reported K_d for 300 K.

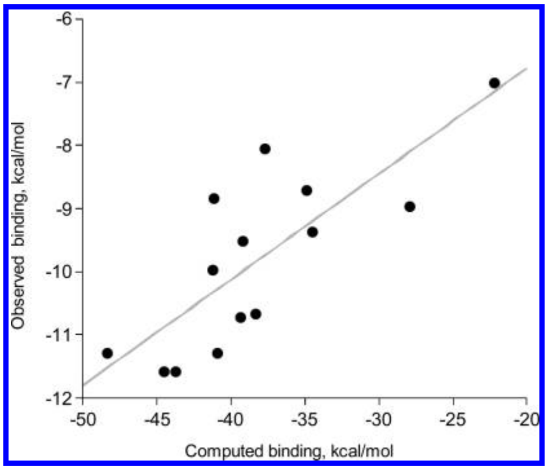

Figure 9. Computed vs experimental binding free energies of p38 ligands.

Table 4. Number of Poses Assembled in the Binding Pocket and in the Reference State, Z_0, for p38 MAP Kinase Ligands

compound[a]	poses assembled	Z_0
16	94 788 702	2.41×10^{034}
40	1 730 915	5.91×10^{041}
46	86 042 060	5.64×10^{033}
65	592 008	1.96×10^{029}
66	6 394 926	8.94×10^{037}
67	80 203 694	1.69×10^{035}
69	100 150 584	1.51×10^{034}
70	33 664 192	1.51×10^{032}
71	23 176 137	6.65×10^{032}
72	27 470 466	6.75×10^{034}
73	19 307 674	1.33×10^{036}
74	3 049 219	5.75×10^{034}
75	20 220 095	1.54×10^{036}
76	3 457 469	2.96×10^{034}
77	1 571 951	5.28×10^{041}

[a] Compound numbering from Regan et al.[45]

the T4 ligands, the use of conformational energies improves the fit to the experimental data; without the correction the r-squared was 0.55, the standard error of prediction 0.98 kcal/mol.

Table 4 provides the number of poses of the whole molecule assembled in the binding pocket and Z_0, the number of poses in the reference state, for each molecule. The number of poses in the binding pocket is very large; compound 69 has over 100 million poses. A superimposition of the 1 000 lowest energy poses of compound 65 is shown in Figure 10. The pyrazole and urea fit tightly into the binding pocket; the assembly process, however, sampled numerous examples of both the expected pose and a higher energy pose of the phenyl pyrazole moiety, separated by a 180° rotation about the bond to the urea. This demonstrates the increased conformational sampling and, therefore, rich information on putative binding modes, inherent in this fragment-based assembly process; dynamics based methods are unlikely to sample both conformations in time scales typically accessible to computation.

The Z_0 values represent $\sim 10^{10}$ translational and rotational poses for the starting fragment, which are simply counted, multiplied by $\sim 10^9$ rotational poses for each bond formed in building the molecule. Despite these large numbers, a fragment-based method in which conformational freedom is accounted for at the point of bond formation is a conceptually simple way to handle conformationally flexible molecules.

DISCUSSION

Computation of binding free energies is straightforward from systematically sampled data. Similar approaches to integrate free energy data from poses have been reported but use a limited and nonuniform selection of poses generated by docking programs.[50,51] The ability to sample energetically relevant poses at very high resolution frees the method from dependence on the choice of an initial pose. This is one of the strengths of this method as compared to perturbation methods, which generally compute the binding free energy only for a selected pose. Figure 11 shows the lowest energy poses of 13 observed binding modes of benzofuran on the T4 lysozyme, superimposed on the pose observed in the crystal structure. All 13 binding modes are sampled and contribute to the computed free energy. Figure 12 shows the full distribution of poses for the lowest energy binding mode. The integration over all poses provides a total binding free

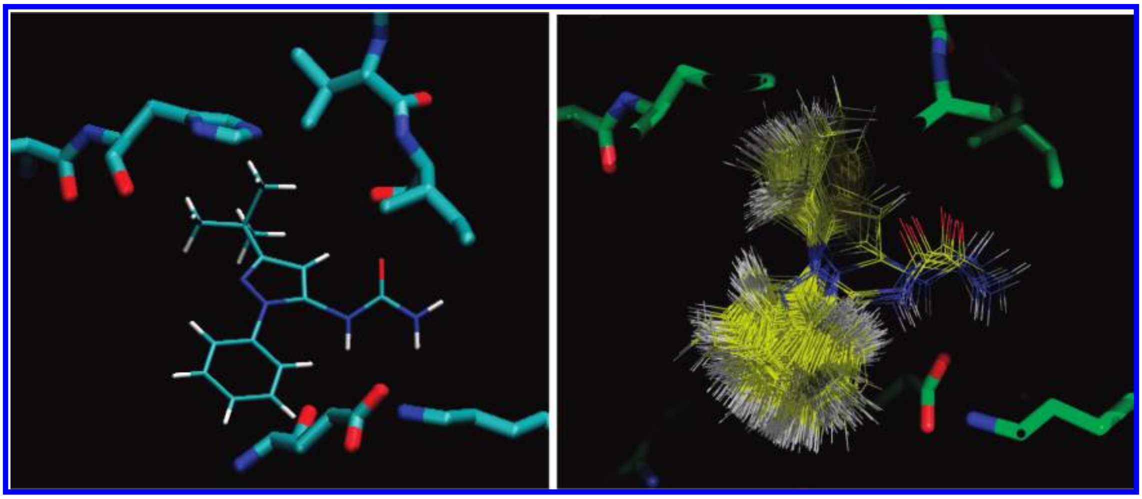

Figure 10. The lowest (left) and 1 000 lowest (right) energy poses of p38 MAP kinase ligand compound 65 (R = H). This demonstrates the sampling of two binding modes, one low-energy mode with the phenyl at the lower left and another high-energy mode with the phenyl at the upper left. Both sets of poses are included in the free energy integration.

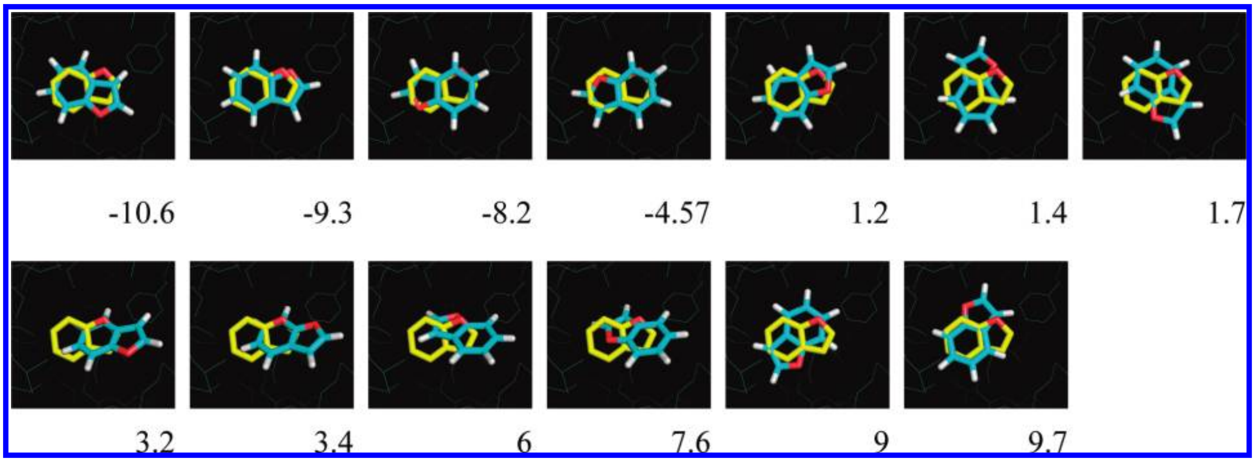

Figure 11. Thirteen computed binding modes of benzofuran for T4 lysozyme L99A and their free energies of binding overlaid on the crystallographically observed pose in yellow. Energies are in kcal/mol.

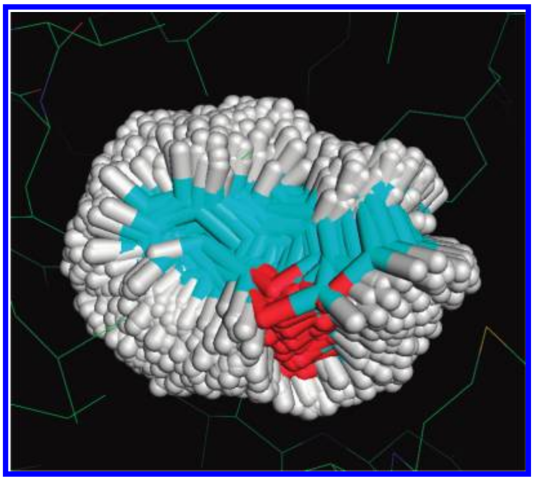

Figure 12. Computed ensemble for the lowest free energy binding mode of benzofuran on T4 lysozyme L99A.

energy of −11.0 kcal/mol, as compared to the −10.6 kcal/mol free energy of the minimum energy pose. The binding mode observed in the crystal corresponds to the mode with the second-lowest computed energy, −9.3 kcal/mol. The crystallographers, however, reported that there was some ambiguity in the assignment of the crystallographic pose. A second advantage to systematic sampling is that there is no need to correct free energies for the symmetry of the fragments. Since all poses are sampled without regard to symmetry, the redundancy of poses due to symmetry transformations is compensated by the same redundancy in the unbound reference state, hence, it does not affect the free energy. This causes the sampling of symmetric molecules, such as benzene, to be less efficient than if symmetry operators were used but has the advantage that no special treatments need to be applied for fragments with and without symmetry.

The limiting entropy for the sampling conditions of T4 lysozyme is about −12.8 kcal/mol. The tightest-fitting single-fragment ligand is the decoy ligand naphthalene. The naphthalene ligand has a computed entropy of −11.2 kcal/mol (Table 2), which is close to the limit. Indeed, only 2 897 poses fit in the binding site with energies less than −5 kcal/mol out of 2.0 × 10^9 poses sampled.

The large ligands in Table 1, that are assembled from multiple fragments, show larger entropy losses, for example, −13.8 kcal/mol for 2-ethyltoluene. These larger entropy losses arise from the introduction of conformational freedom on bond formation and the subsequent loss of conformational flexibility upon binding. The effect of this conformational flexibility is to increase the number of poses that must be sampled in the binding site, beyond the 2.0 × 10^9 poses that are sampled for the T4 lysozyme site, for example. Since the number of poses sampled must increase combinatorially with the number of rotatable bonds, the computational cost of sampling flexible molecules can rapidly become prohibitive. One of the reasons for using a fragment-based free energy calculation, therefore, is that the relatively high entropies associated with complex molecules can be more efficiently sampled by breaking them into smaller fragments and addressing the conformational combinatorics in the fragment assembly stage.

The magnitudes of the entropies of the p38 ligands, built from four fragments, are similar to those reported by Gilson for the entropy loss of Amprenavir on binding to HIV protease.[52] In that case, the total entropy loss is computed to be −26.4 kcal/mol, as compared to our highest entropy of −19.6 kcal/mol for *iso*-butyl benzene in T4 and the average of −20.5 kcal/mol for the p38 compounds. For both T4 and p38, the computed binding free energy is a better predictor of binding than the enthalpy alone, which has prediction errors for the two proteins of 1.78 and 0.93 kcal/mol, respectively, suggesting that the computed entropy contribution is improving the prediction over the computed enthalpy alone.

The T4 lysozyme system is a good test case because the protein is relatively rigid, and the binding site is completely divorced from solvent contact.[41] However, the small range of measured binding values (ranging from −4.6 to −6.7 kcal/mol) handicaps efforts, since the accuracy of force field calculations may be no better than 1 kcal/mol.[24,25,4] The variance of 0.83 kcal/mol of measured binding energies of the series combined with a force field error of 1 kcal/mol would result in an r-squared of about 0.5. In this study, the r-squared is nearly exactly 0.5 for the rigid fragments and 0.74 for all compounds, suggesting that the sampling may be providing results near the limit of the molecular mechanics method. The standard error of prediction for the T4 ligands is 1.2 kcal/mol. The error in the prediction of p38 binders is somewhat lower at 0.88 kcal/mol, while linear interaction energy calculations on the same set of p38 compounds resulted in a standard error of 0.77 kcal/mol.[53] These results for p38 add further credence to the evidence that force field methods have an accuracy threshold near 1 kcal/mol.

The ability to combinatorially score a very large number of molecules with a relatively few free energy computations is an advantage of the fragment-based approach. The assembly process can construct and score thousands or millions of molecules by combining a relatively small set of fragments. The combination of fragments involves the approximation that the atomic charges of the fragments are similar to the charges that the atoms will have in the assembled molecule. Since a hydrogen atom is “removed” from each fragment to model the creation of a chemical bond, the energy contributions of these hydrogens are included in the free energy estimate for the assembled molecule. If either of these atoms is polar, then the energy contribution may be significant. The error created by this assumption is ameliorated by the choice of fragments and connection points. In our fragment library, polar hydrogen atoms that will be connection points are replaced with methyl groups, and the entire methyl group is removed when the fragment is joined at that point.

Since the method described integrates the free energy, including the entropy of combining fragments, the resulting energy is not the simple sum of the fragment free energies. For example, the binding energies of the individual fragments comprising *n*-butyl benzene sum to −18.8 kcal/mol (benzene at −9.4 and twice ethane at −4.7 kcal/mol each), while the reintegrated binding free energy is only −15.0 kcal/mol. The difference arises from the favorable change in entropy associated with combining fragments into molecules and a generally unfavorable change in entropy due to the restricted motion of a molecule in the binding site, compared to the freedom of its smaller component fragments. The inclusion of the conformational and interaction energies among fragments when they are connected further improves the result, as shown by the comparison of Figures 8 and 7. In comparison, perturbation methods normally compute the free energy of the entire molecule at once and sample only in the vicinity of the starting pose. While perturbation methods do not require the complex integration of fragment entropies, they may be susceptible to error by lack of inclusion of the lower energy conformations.[26] Consistent with our results for binding of *n*-butyl and *iso*-butyl benzene to T4 lysozyme L99A, conformational energies were estimated by other workers to contribute about 5 kcal/mol to the free energy of binding a series of ligands to HIV reverse transcriptase.[45]

The rigid protein and fragment approximation allows fast estimation of binding for a large set of compounds but is also a potential weak point. The assumed rigidity does not represent the motion of the protein to accommodate ligands, while this flexibility is generally included in perturbation methods. In this study, the protein form with the largest binding pocket was used to allow all of the ligands to fit into the site. The use of building tolerances for assembling fragments allows some of the required protein flexibility to be transferred to the ligand; the ligand can adopt nonideal angles and lengths to fit into the rigid protein structure. As long as the energy required to rearrange the protein for different ligands is small, or close to *kT*, the approximation of rigid protein is acceptable, and we find that in many cases the computed binding for the larger pocket size represents the observed rank order of binding for a series of ligands fairly well. There are certainly cases where this approximation will cause loss of accuracy. In those situations, several protein forms may be used for a series of simulations, but in these cases, the different free energies of the different protein forms may contribute significantly to the binding free energy.

The treatment of solvation energies is also an approximation to this method. The fragment gas-in-aqueous-phase free energy is applied as a correction at the fragment level, but the desolvation of the protein is not included in the calculation of free energy. We have computed the solvation energy of the T4 and p38 (for the R-substituent) binding pockets, by treating water as a fragment in grand canonical simulations, and found that the affinity of water for these sites is very low, 1.2 kcal/mol for the T4 binding site. This calculation is consistent with experimental studies of water in the T4 lysozyme L99A site, which show that at 1 atm the electron density is consistent with an average of 1.5 water molecules in the site at crystallographic temperatures.[54,55] It is not probable, therefore, that solvation of the protein is a major factor in the correlations for these specific examples. There are obviously other cases, however, where there is strong solvation of the protein, which must be accounted for.

Although the approximations used herein for protein flexibility and protein solvation were suitable for rank-ordering ligand affinities, the slopes of computed vs observed free energies are 1/2.9 for the T4 ligands and 1/3.7 for p38, suggesting that there are similar sources of systematic error in both cases. Since protein solvation is weak in this case and the major binding interactions for the ligands we are simulating are hydrophobic, it is unlikely that the source of errors is a function of solvation or electrostatic terms in the force field. It is possible that the systematic error arises from the van der Waals interaction term of the force field, but we have no evidence to confirm this. Our current hypothesis for the source of the systematic error on the slope is the approximation of using a single protein structure for all simulations. To the extent that the smallest fragments tend to have the weakest affinities, using a single binding pocket for all fragments will overestimate the size of the pocket for the small fragments and, therefore, underestimate their binding interactions, particularly for van der Waals interactions. This would have the effect of increasing the slope of the correlation.

The principle that the simple electrostatic energy is incorrect in an absolute measure but still related to binding energy is supported by the success of the linear interaction energy methods.[56−58] These methods derive coefficients to computed values by comparison to experimental data in order to improve the prediction of binding free energy, implying that there is an underlying linear correlation. In this study, no modeling of coefficients is required to compute relative free energies. Current work with the systematic sampling free energy method is focusing on improved treatments of solvation and protein energies.

Deng and Roux have recently reported free energy calculations on the T4 lysozyme L99A system.[4] Their method, using a sophisticated treatment of solvation as well as a flexible protein model, provides a binding free energy very close to the experimental value. This demonstrates that a complete potential model can give results close to the experimental binding. However, such a complete treatment is computationally intensive, so that the binding energy is computed for only a single pose, and the results of the dynamics-based method are dependent on the choice of starting pose and restraint parameters. The binding free energy of benzene to T4 was also studied in detail by Hermans and Wang, using slow growth thermodynamic integration.[59] While that work did not compute free energies of other T4 ligands, they did report a binding free energy of −7 to −9 kcal/mol for benzene, which is comparable to our value of −9.4 kcal/mol and the −9.6 kcal/mol provided by grand canonical sampling.[6] In addition, the reported mean benzene binding enthalpy of −15.2 kcal/mol compares well to the enthalpy of −16.0 in this work and the −16.0 kcal/mol computed with grand canonical sampling. The results from our computations are, therefore, in general agreement with those of previous researchers. No method that is fast enough for the purposes of drug design, including the current method, has yet implemented a complete binding free energy computation, including ligand−protein interaction free energies, ligand conformational free energies, protein structural free energies, and solvation free energies. We find, however, that the systematic sampling method will consistently predict the rank order of ligand affinities within a standard error of prediction near 1 kcal/mol, provided that tightly bound water is identified and accounted for.[22]

CONCLUSIONS

The systematic sampling of binding energies provides an effective sampling means for computing of protein−ligand binding free energies. The rigor of sampling and reintegration of free energies, within the defined translation and rotational parameters, allows the method to be independent of the selected starting poses and largely independent of the fragment combination used to assemble the molecules. The method of connecting fragments and reintegrating the free energy is among the first to address a recognized issue in fragment-based drug design − the weak approximation of simply adding the contribution of each fragment − when estimating the binding of complex molecules.

The method described herein does not deliver absolute free energies, possibly due to the use of a simplistic treatment of protein structures; however, it does provide values that are well correlated with observed binding free energies. The method can readily distinguish known nonbinding compounds from known binding compounds on the basis of the computed binding free energies.

ACKNOWLEDGMENT

We thank Jennifer Ludington, Ted Fujimoto, Eric Springman, Tony Klon, Zenon Konteatis, Jinming Zou, Igor Shkurko, and Gergely Tóth for assistance and discussions.

Supporting Information Available: The alogrithm to compute set S_1 and protein modeling details. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Waller, C. L. Recent Advances In Molecular Diversity. *J. Comp.−Aided Mol. Des.* **2000**, *16*, 299–300.
(2) Ekins, S.; Boulanger, B.; Swaan, P. W.; Hupcey, M. A. Z. Towards a new age of virtual ADME/TOX and multidimensional drug discovery. *J. Comp.−Aided Mol. Des.* **2000**, *16*, 381–401.
(3) Schnecke, V.; Bostrom, J. Computational Chemistry-Driven Decision Making in Lead Generation. *Drug Discov. Today* **2006**, *11*, 43–50.
(4) Tickle, I.; Sharff, A.; Vinkovic, M.; Yon, J.; Jhoti, H. High-Throughput Protein Crystallography And Drug Discovery. *Chem. Soc. Rev.* **2004**, *33*, 558–565.
(5) Hann, M. M.; Oprea, T. I. Pursuing The Lead Likeness Concept in Pharmaceutical Research. *Curr. Opin. Chem. Biol.* **2004**, *8*, 255–263.
(6) Morelli, X.; Rigby, A. C. Acceleration Of The Drug Discovery Process: A Combinatorial Approach Using NMR Spectroscopy and Virtual Screening. *Curr. Comput.−Aided Drug Des.* **2007**, *3*, 33–49.
(7) Reynolds, C. H.; Tounge, B. A.; Bembenek, S. D. Ligand binding efficiency: trends, physical basis, and implications. *J. Med. Chem.* **2008**, *51*, 2432–2438.
(8) Jencks, W. P. On the Attribution and Additivity of Binding Energies. *Proc. Natl. Acad. Sci., U.S.A.* **1981**, *78*, 4046–4050.
(9) Lu, B.; Wong, C. F. Direct Estimation of entropy Loss due to Reduced Translational and Rotational Motions upon Molecular Binding. *Biopolymers* **2005**, *79*, 277–285.
(10) Murray, C. W.; Verdonk, M. L. The consequences of translational and rotational entropy loss by small molecules on binding to proteins. *J. Comp.−Aided Mol. Des.* **2002**, *16*, 741–753.
(11) 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*, 5912–31.
(12) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. The Statistical-Thermodynamic Basis for Computation of Binding Affinities: A Critical Review. *Biophys. J.* **1997**, *72*, 1047–1069.

(13) Gilson, M. K.; Zhou, H.-X. Calculation of Protein-Ligand Binding Affinities. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.

(14) Deng, Y.; Benot Roux, B. Computations of Standard Binding Free Energies with Molecular Dynamics Simulations. *J. Phys. Chem. B* **2009**, *113*, 2234–2246.

(15) Shirts, M.; Mobley, D.; Chodera, J. Alchemical free energy calculations: Ready for prime time. *Annu. Rep. Comput. Chem.* **2007**, *3*, 41–59.

(16) Zwanzig, R. W. High-Temperature Equation Of State By a Perturbation Method. *J. Chem. Phys.* **1954**, *22*, 1420–1426.

(17) Pranata, J.; Jorgensen, W. L. Monte-Carlo Simulations Yield Absolute Free-Energies of Binding for Guanine-Cytosine And Adenine Uracil Base Pairs in Chloroform. *Tetrahedron* **1991**, *47*, 2491–2501.

(18) Hamelberg, D.; McCammon, J. A. Standard free-energy of Releasing a Localized Water Molecule from the Binding Pockets of Proteins: Double-Decoupling Method. *J. Am. Chem. Soc.* **2004**, *126*, 7683–7689.

(19) Roux, B.; Nina, M.; Pomes, R.; Smith, J. C. Thermodynamic Stability Of Water Molecules In The Bacteriorhodopsin Proton Channel: A Molecular Dynamics Free Energy Perturbation Study. *Biophys. J.* **1996**, *71*, 670–681.

(20) Hermans, J.; Subramaniam, S. The Free Energy of Xenon Binding to Myoglobin From Molecular Dynamics Simulation. *Isr. J. Chem.* **1988**, *27*, 225–227.

(21) Wang, J.; Morin, P.; Wang, W.; Kollman, P. A. Use of MM-PBSA in Reproducing the Binding free-energies to HIV-1 RT of TIBO Derivatives and Predicting the Binding Mode to HIV-1 RT of Efavirenz by Docking and MM-PBSA. *J. Am. Chem. Soc.* **2001**, *123*, 5221–5230.

(22) Clark, M.; Guarnieri, F.; Shkurko, I.; Wiseman, J. Grand Canonical Monte Carlo Simulation of Ligand-Protein Binding. *J. Chem. Inf. Model.* **2006**, *46*, 231–242.

(23) Guarnieri, F. Computational Protein Probing to Identify Binding Sites. U. S. Patent 6,735,530, May 11, 2004.

(24) Weis, A.; Katebzadeh, K.; Soderhjelm, P.; Nilsson, I.; Ryde, U. Ligand Affinities Predicted with the MM/PBSA Method: Dependence on the Simulation Method and the Force Field. *J. Med. Chem.* **2006**, *49*, 6596–6606.

(25) 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.

(26) Tirado-Rives, J.; Jorgensen, W. L. Contribution of Conformer Focusing to the Uncertainty in Predicting free-energies for Protein-Ligand Binding. *J. Med. Chem.* **2006**, *49*, 5880–5884.

(27) Mobley, D. L.; Chodera, J. D.; Dill, K. A. On the Use of Orientational Restraints and Symmetry Corrections in Alchemical Free Energy Calculations. *J. Chem. Phys.* **2007**, *125*, 084902.

(28) Mobely, D. L.; Graves, A. P.; Chodera, J. P.; McReynolds, A. C.; Shoichet, B. K.; Dill, K. A. Predicting Absolute Ligand Binding Free Energies to a Simple Model Site. *J. Mol. Biol.* **2007**, *371*, 1118–1134.

(29) Matsumura, M.; Becktel, W. J.; Levitt, M.; Matthews, B. W. Stabilization of Phage T4 Lysozyme by Engineered Disulfide Bonds. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 6562–6566.

(30) Gray, T. M.; Arnoys, E. J.; Blankespoor, S.; Born, T.; Jagar, R.; Everman, R.; Plowman, D.; Stair, A.; Zhang, D. Destabilizing Effect of Proline Substitutions in Two Helical Regions of T4 Lysozyme: Leucine 66 To Proline and Leucine 91 to Proline. *Protein Sci.* **1996**, *4*, 742–751.

(31) Faber, H. R.; Matthews, B. W. A Mutant T4 Lysozyme Displays Five Different Crystal Conformations. *Nature* **1990**, *348*, 263–266.

(32) de Groot, B. L.; Hayward, S.; van Aalten, D. M. F.; Amadei, A.; Berendsen, H. J. C. Domain Motions in Bacteriophage T4 Lysozyme; a Comparison between Molecular Dynamics and Crystallographic Data. *Proteins: Struct., Funct., Gen.* **1998**, *31*, 116–127.

(33) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of Protein-Ligand Interactions. Docking and Scoring: Successes and Gaps. *J. Med. Chem.* **2006**, *49*, 5851–5855.

(34) Ferrarri, A. M.; Wei, B. Q.; Costantino, L.; Shoichet, B. K. Soft Docking and Multiple Receptor Conformations in Virtual Screening. *J. Med. Chem.* **2004**, *47*, 5076–5084.

(35) Graves, A. P.; Brenk, R.; Shoichet, B. K. Decoys for Docking. *J. Med. Chem.* **2005**, *48*, 3714–3728.

(36) Deng, Y. Roux, Benoît Calculation of Standard Binding free-energies: Aromatic Molecules in the T4 Lysozyme L99A Mutant. *J. Chem. Theory Comput.* **2006**, *2*, 1255–1273.

(37) Dinarello, O. A. Inflammatory Cytokines: Interleukin 1 and Tumor Necrosis Factor as an Effector Molecules in Autoimmune Diseases. *Curr. Opin. Immunol.* **1991**, *3*, 941–948.

(38) Karney, C. F. F. Quaternions in Molecular Modeling. *J. Mol. Graphics Modell.* **2006**, *25*, 595–604.

(39) Cornell, W. D.; Cieplak, P. I.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D.; 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.

(40) Breneman, C.; Wiberg, K. B. Determining atom-centered monopoles from molecular electrostatic potentials. The Need for High Sampling Density in Formamide Conformational Analysis. *J. Comput. Chem.* **1990**, *11*, 361–373.

(41) Morton, A.; Baase, W. A.; Matthews, B. W. Energetic Origins of Specificity of Ligand Binding in an Interior Nonpolar Cavity of T4 Lysozyme. *Biochemistry* **1995**, *34*, 8564–8575.

(42) Morton, A.; Matthews, B. A. Specificity of Ligand Binding in a Buried Nonpolar Cavity of T4 Lysozyme: Linkage of Dynamics and Structural Plasticity. *Biochemistry* **1995**, *34*, 8576–8588.

(43) *Macromodel 8.6*; Schrödinger Inc.: New York, NY, 2004.

(44) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck; A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B. G.; Chen, W.; Wong, M. W.; Andres, J. L.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98* (Revision A.9); Gaussian, Inc.: Pittsburgh PA, 1998.

(45) Regan, J.; Breitflelder, S.; et al. Pyrazole Urea-Based inhibitors of p38 MAP Kinase: From Lead Compound to Clinical Candidate. *J. Med. Chem.* **2002**, *45*, 2994–3008.

(46) Pargellis, C.; Tong, L.; Churchill, L.; Cirillo, P. F.; Gilmore, T.; Graham, A. G.; Grob, P. M.; Hickey, E. R.; Moss, N.; Pav, S.; Regan, J. Inhibition Of P38 MAP Kinase by Utilizing a Novel Allosteric Binding Site. *Nat. Struct. Biol.* **2002**, *9*, 268–272.

(47) Wang, Z.; Canagarajah, B. J.; Boehm, J. C.; Kassisa, S.; Cobb, M. H.; Young, P. R.; Abdel-Meguid, S.; Adams, J. L.; Goldsmith, E. J. Structural Basis of Inhibitor Selectivity In MAP Kinases. *Structure* **1998**, *6*, 1117–1128.

(48) Mobley, D. L.; Chodera, J. D.; Dill, K. A. Confine-and-Release Method: Obtaining Correct Binding Free-Energies in The Presence of Conformational Change. *J. Chem. Theory Comput.* **2007**, *3*, 1231–1235.

(49) Shoichet group web page. http://shoichetlab.compbio.ucsf.edu (accessed March 2009).

(50) Ruvinsky, A. M. Calculations of protein-ligand binding entropy of relative and overall molecular motions. *J. Comput.−Aided Mol. Des.* **2007**, *21*, 361–370.

(51) Michel, J.; Essex, J. W. Hit Identification and Binding Mode Predictions by Rigorous free-energy Simulations. *J. Med. Chem.* **2008**, *51*, 6654–6664.

(52) Chang, C.-E.; Chen, W.; Gilson, M. K. Ligand configurational entropy and protein binding. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1534–1539.

(53) Tokinaga, Y.; Jorgensen, W. L. General Model for Estimation of the Inhibition of Protein Kinases Using Monte Carlo Simulations. *J. Med. Chem.* **2004**, *47*, 2534–2549.

(54) Collins, M. D.; Hummer, G.; Quillin, M. L.; Matthews, B. W.; Gruner, S. M. Cooperative Water Filling of a Nonpolar Protein Cavity Observed by High-Pressure Crystallography and Simulation. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 16668–16671.

(55) Liu, L.; Quillin, M. L.; Matthews, B. W. Use of experimental crystallographic phases to examine the hydration of polar and nonpolar cavities in T4 lysozyme. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 14406–14411.

(56) Åqvist, J.; Marelius, J. The Linear Interaction Energy Method for Predicting Ligand Binding free-energies. *Comb. Chem. High Throughput Screening* **2001**, *4*, 613–626.

(57) Rizzo, R. C.; Udier-Blagovic, M.; Wang, D.-P.; Watkins, E. K.; Kroeger Smith, M. B.; Smith, R. H., Jr.; Tirado-Rives, J.; Jorgensen, W. L. Estimation of Binding Affinities for HEPT and Nevirapine Analogues with HIV-1 Reverse Transcriptase via Monte Carlo Simulations. *J. Med. Chem.* **2002**, *45*, 2970–2987.

(58) Pierce, A. C.; Jorgensen, W. L. Estimation of Binding Affinities for Selective Thrombin Inhibitors via Monte Carlo Simulations. *J. Med. Chem.* **2001**, *44*, 1043–1050.

(59) Hermans, J.; Wang, L. Inclusion of loss of translational and rotational freedom in theoretical estimates of free-energies of binding. Application to a complex of benzene and mutant T4 lysozyme. *J. Am. Chem. Soc.* **1997**, *119*, 2707–2714.

CI900132R