

Modeling Protein–Ligand Binding by Mining Minima

Wei Chen,[†] Michael K. Gilson,^{*,†} Simon P. Webb, and Michael J. Potter^{*}

VeraChem LLC, P.O. Box 2206, Germantown, Maryland 20875-2206

Received May 10, 2010

Abstract: We present the first application of the mining minima algorithm to protein–small molecule binding. This end-point approach uses an empirical force field and implicit solvent models, treats the protein binding site as fully flexible, and estimates free energies as sums over local energy wells. The calculations are found to yield encouraging agreement with experimental results for three sets of HIV-1 protease inhibitors and a set of phosphodiesterase 10a inhibitors. The contributions of various aspects of the model to its accuracy are examined, and the Poisson–Boltzmann correction is found to be the most critical. Interestingly, the computed changes in configurational entropy upon binding fall roughly along the same entropy–energy correlation previously observed for smaller host–guest systems. Strengths and weaknesses of the method are discussed, as are the prospects for enhancing accuracy and speed.

1. Introduction

The appeal and promise of structure-based drug design has grown in recent decades as computer power has increased and as the three-dimensional structures of proteins have become easier to solve by crystallographic and NMR methods (e.g., refs 1–6). Indeed, advances in protein crystallography are now extending the reach of structure-based approaches into the medically critical yet previously inaccessible realm of G-protein coupled receptors.^{7–11} Computational tools for the selection or design of compounds to bind a targeted protein also have advanced substantially in recent years (e.g., refs 12–17), but there is still a need for more reliable methods of ranking candidate ligands of a targeted protein according to affinity.^{18,19}

Approaches to modeling protein–small molecule interactions may be viewed, very broadly, as falling along a spectrum of complexity. At one end lie the simplest and most efficient docking methods (e.g., refs 20–28), which seek to identify the single most stable conformation of a protein–ligand complex and estimate binding affinity with a scoring function that is often conceptualized as a sum of free energy contributions from, for example, hydrogen bonding, the hydrophobic effect, and Coulomb interactions. (Recent

progress in the use of relatively detailed electronic structure calculations as part of protein–ligand scoring functions also deserves mention (e.g., refs 29–36).) These methods offer the speed required to rapidly screen large compound databases for promising candidate ligands but typically neglect or, arguably, oversimplify significant free energy contributions, such as the energetic and entropic consequences of ligand preorganization or the lack thereof. At the other end of the spectrum lie free energy pathway methods, like free energy perturbation and thermodynamic integration, which employ relatively detailed Monte Carlo or molecular dynamics methods, typically using explicit solvent models, to compute the absolute or relative work of binding for candidate ligands (e.g., refs 37–47). Such approaches have the potential to capture much of the relevant physics but still tend to be too computationally demanding for routine use in drug-design projects. Toward the middle of the spectrum of complexity lie so-called “end-point” free energy methods, such as MMPBSA^{48–51} and LIE.^{52–55} These aim to provide greater physical detail than docking and scoring methods by accounting for both the bound and unbound states of the protein–ligand system, while potentially avoiding the computational costs of free-energy simulations because they do not involve computing the work of a multistep binding process or “alchemical”³⁷ ligand change.

The present study focuses on one of a class of end-point methods,⁵⁶ including mining minima (here M2),^{57–59} and MINTA,⁶⁰ which approximate the thermodynamics of bind-

* To whom correspondence should be addressed. Phone: (240) 686-0565. Fax: (240) 686-0564. E-mail: gilson@verachem.com (M.K.G.); potterm@verachem.com (M.J.P.).

[†] These authors contributed equally to this work.

ing by identifying a manageable small set of conformations (local energy minima) of the free and bound protein and ligand. The free energy associated with each local energy minimum of the ligand, receptor or complex, is evaluated on the basis of its depth and width, and the contributions of the energy wells are combined to yield an estimate of the overall free energy. A potential merit of this approach is that one may employ conformational search approaches that aggressively cross energy barriers to identify the stable conformations, rather than having to wait for barriers to be crossed through thermal motion, as in typical simulation-based methods. On the other hand, in order to avoid a combinatorial explosion of conformations, such methods must limit the number of explicit degrees of freedom by using an implicit solvent model and, for large receptors, treating part of the system as restrained or rigid. Applications of such models to binding in host–guest systems have shown encouraging agreement with experimental results,^{57,59–62} suggesting that this approach, used with current energy models, captures much of the relevant physical chemistry of molecular recognition. The host–guest studies also have yielded unexpected insights into the role of configurational entropy in binding, providing evidence that these entropy contributions are commensurate with more familiar energy terms and may differ enough across ligands to significantly affect their affinity rankings. It is thus of considerable interest to apply this approach to protein–small molecule binding.

We have now developed an implementation of the M2 algorithm, VM2, that is suitable for protein–ligand affinity calculations. The software includes more aggressive search algorithms than those used for the prior host–guest calculations and allows part of the target protein to be held rigid while a user-defined binding site region, including a ligand and any number of additional solvent molecules or cofactors, is treated as flexible. This paper describes the algorithm and characterizes its performance for three groups of compounds binding two very different proteins, the human immunodeficiency virus 1 protease (HIVP) and phosphodiesterase 10a (PDE 10a).

2. Methods

2.1. Application of Mining Minima to Protein Ligand Binding. *2.1.1. Overview.* As previously described,^{57,59} the second-generation mining minima method, M2, computes the standard free energy (or, more properly, the standard chemical potential) of a bound complex, AB, and the free molecules, A and B, and obtains the standard binding free energy as the difference:

$$\Delta G^\circ = G_{AB}^\circ - G_A^\circ - G_B^\circ \quad (1)$$

M2 is thus an “end-point” method.¹² In brief, the free energy, G_X , of a molecule or complex $X = A, B$, or AB is estimated by summing local configuration integral contributions z_i from distinct energy wells (conformations) i :

$$G_X^\circ \approx -RT \ln \sum_i z_i \quad (2)$$

Each local integral, z_i , is obtained by computing the Hessian matrix with respect to bond-angle-torsion coordinates at the

corresponding local energy minimum and applying an enhanced harmonic approximation which accounts for anharmonicity along those eigenvectors of the Hessian having eigenvalues up to 2 kcal/mol/xy, where x and y are either Angstroms or radians, depending on the coordinates involved. Local energy minima are found with an aggressive conformational search which, unlike typical molecular dynamics simulations, can readily cross energy barriers (see below). The energy model comprises an empirical force field and an implicit solvent model.

The present study uses an implementation of M2 that is suitable for protein–ligand modeling. In particular, it allows part of the receptor (the “real set”) to be held rigid while treating only a user-defined binding-site region (the “live set”) as flexible. The rigid part of the receptor acts as a framework that holds the flexible binding site region in the right overall conformation, while the flexibility of the binding site allows its conformation to adapt to different ligands. The present implementation also affords additional, more aggressive, conformational search methods and permits the complex to include more than one bound molecule; for example, it can include a drug-like ligand and one or more explicit water molecules. A ligand or solvent molecule can, optionally, be restrained in a spherical region near a desired location by applying a flat-bottomed energy well to one of its atoms:

$$E_{\text{restraint}} = k \frac{(r - r_0)^p}{R_{\text{restraint}}^p} \quad (3)$$

Here, k is the force constant of the restraint, r is the location of the restrained atom, r_0 is the center of the flat-bottomed well, in coordinates rooted in the rigid part of the receptor, $R_{\text{restraint}}$ is the radius of the restraint region, and p is a parameter controlling the “hardness” of the energy wall around the energy well. A value of $p = 12$ was used in all calculations presented here.

2.1.2. Conformational Search. The conformational search method is a variant of the previously described Tork algorithm,⁶³ which in turn is inspired by the low mode search method.^{64,65} The search typically starts with a single initial conformation, although more than one can be used. The conformation is energy-minimized to a local energy minimum, and the second-derivative (Hessian) matrix of the energy in Cartesian coordinates is computed, transformed into the Hessian for bond-angle-torsion coordinates, as previously described,⁶³ reduced by elimination of rows and columns corresponding to bond-stretch and bond-angle coordinates, and diagonalized. A “distortion mode” is then defined by choosing a normalized eigenvector e_i with a low eigenvalue v_i , or a linear combination of two normalized eigenvectors e_i and e_j with low eigenvalues v_i and v_j . The mode is then modified by zeroing any eigenvector components less than 0.1. The system is distorted along this distortion mode, and then 10 steps of quasi-Newton energy minimization are carried out with the distorted atoms held fixed and the rest of the system free. If the final energy is less than a threshold of E_t and no torsional change has exceeded a user-defined threshold of ϕ_t , then the process is

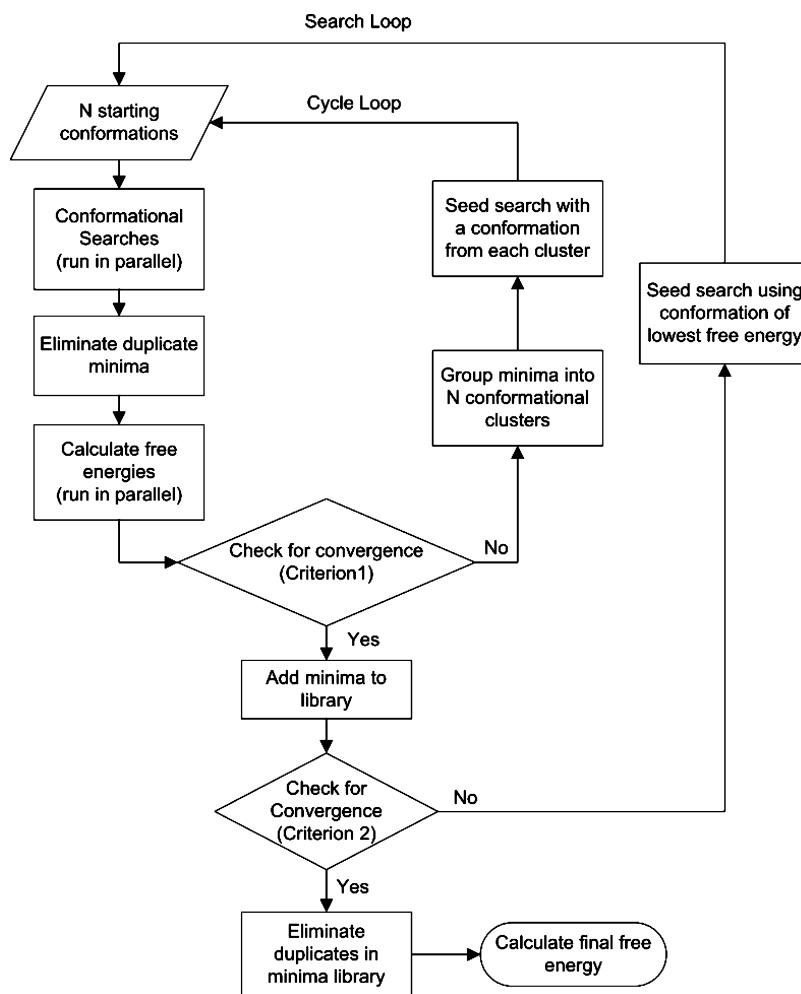


Figure 1. High-level flow-chart of the mining minima procedure, showing nested iterative loops over cycles and searches, each with its own convergence criterion, as detailed in the text.

repeated. When this iterative process crosses either the energy threshold or the torsional threshold, the distorted conformation is fully energy-minimized to a local energy minimum. Here, the 100 eigenvectors with the lowest eigenvalues were considered “low”, the energy threshold, E_t , was set to 2000 kcal/mol, and the torsional threshold, ϕ_t , was set to 180°. For the pairwise distortions, the contribution of each eigenvector to the combined distortion vector, e_{combo} , was weighted according to the reciprocal of its eigenvalue:

$$e_{\text{combo}} = \frac{e_i/v_i + e_j/v_j}{1/v_i + 1/v_j} \quad (4)$$

The present software implementation also includes two aggressive search methods. In one, only the ligand is moved stepwise along its soft modes, and after each step, the live atoms of the protein are energy-minimized for several steps so that they move out of the way of the moving ligand. This process is iterated until an energy or displacement is reached, at which point the entire system is energy-minimized. The second aggressive search method is the same, except that only the protein is moved stepwise along its soft modes, while the ligand is relaxed in response. These more aggressive methods are most appropriate for cases where the binding modes are unknown and were not used here because

we wished to leave the ligands’ scaffolds close to their positions in the available crystallographic structures.

2.1.3. Monitoring Free Energy Convergence. The free energy calculation for the protein, ligand, or complex begins with a single energy-minimized conformation. As diagrammed in the flowchart in Figure 1, the algorithm then follows a convergent, iterative procedure involving a series of “searches”, each composed of “cycles”, where each search is run to convergence (criterion 1 in the flowchart) and the overall set of searches is also run to convergence (criterion 2 in the flowchart). The first search is carried out as follows. The starting conformation is used to seed a Tork conformational search (see prior section) and thereby generate a set of new conformations. Any duplicate conformations are removed from the list, and the configuration integral z_i is computed for each of the N distinct local energy minima. The free energy is obtained from eq 2, where the sum ranges over the energy minima i found to date.

This completes the first cycle of the first search. All conformations from the current search iteration are then superimposed and grouped into N clusters based upon the atomic root-mean-square distances between conformations, and the lowest free energy conformation of each cluster is used to start a new Tork conformational search. Thus, each cycle after the initial one starts with N , rather than 1, initial

conformations, but the algorithm is otherwise the same. This procedure aims to generate a diversity of starting structures for the new cycle. Duplicates are again removed, with reference to all conformations generated in the current search. Configuration integrals are computed for the remaining distinct conformations, and the free energy is updated to reflect all distinct conformations found in the current cycle. Additional cycles are run in this manner until the last cycle meets criterion 1: it must lower the free energy by less than 0.05 kcal/mol, and the last three cycles must cumulatively lower G by less than 0.5 kcal/mol. At this point, the first search is considered to be converged. A new search then begins from the lowest free energy conformation found in the prior search. Additional searches are run until criterion 2 is met; i.e., the last search must lower the free energy by less than 0.05 kcal/mol. In the present calculations, the number of clusters, N , is set to 4.

2.1.4. Energy Model. The protein's potential energy as a function of conformation was approximated with the CHARMM⁶⁶ united-atom force field. For the ligands, we used VeraChem's enhanced version of the Dreiding⁶⁷ force field for bond-stretch, bond-angle, and torsional terms. The Lennard-Jones terms were automatically assigned to those of chemically related atoms in the CHARMM force field, and partial charges were assigned with the electronegativity equalization method Vcharge.⁶⁸ The topology files for the ligands are provided in the Supporting Information. During the conformational search procedure, solvation is modeled with a generalized Born model (GB).^{69,70} The free energy of each energy minimum is subsequently corrected, as previously described,⁵⁷ by subtracting its GB energy and adding in the Poisson–Boltzmann/surface area (PB/SA) solvation energy, where the Poisson–Boltzmann equation is solved with a fast finite difference code.^{71,72} It is this PB-corrected free energy that is checked to determine whether the new conformation is more stable than prior conformations. Using the corrected free energy helps prevent the search algorithm from finding conformations that are stable according to the GB model but not according to the PB/SA model. The dielectric cavity radius of each atom is set to the value of its Lennard-Jones parameters, i.e., to the distance at which the Lennard-Jones energy of two atoms of this type has its local energy minimum.

2.2. Protein–Ligand Systems and Setups. We applied the method described above to 24 HIV-1 protease inhibitors and 20 phosphodiesterase 10a (PDE 10a) inhibitors. The protease inhibitors are divided into two groups. Group 1 is a heterogeneous collection of drugs and nonclinical compounds with measured binding free energies of -19.2 to -6.6 kcal/mol, as determined by calorimetric and enzymatic K_i measurements cited in the caption of Table 2. Although these data derive from six different research groups, their wide range of binding free energies reduces the quantitative significance of methodologic variations among the groups. Group 1 is subdivided into group 1a compounds (Figure 2), which are known or presumed to bind with a bridging “flap water”, and group 1b compounds (Figure 3), cyclic and azacyclic ureas which bind without a flap water. Group 2 (Figure 4) is a relatively homogeneous set of 10 compounds⁷³

from a project aimed at discovering new protease inhibitors that would be robust to mutational resistance, with binding free energies of -16.7 to -9.3 kcal/mol, on the basis of enzymatic K_i data obtained by a single research group. The PDE 10a inhibitors⁷⁴ possess a common chemical scaffold (Figure 5) and range from -11.5 to -9.2 kcal/mol in binding free energy, also on the basis of K_i data obtained by a single research group.

The evaluation of a method of ranking ligand affinities should include ligand series for which affinity is not strongly correlated with ligand size, so that the arguably trivial tendency of many scoring functions to correlate with ligand size⁷⁵ does not lead to an overly optimistic assessment of the method. Here, the correlation coefficients (R values) of binding free energy with the number of non-hydrogen atoms are 0.75, 0.63, 0.00, and 0.44 for the group 1a, group 1b, and group 2 protease inhibitors and PDE 10a, respectively. We thus have a mix of correlated and uncorrelated series, much as previously observed for a variety of experimental data sets.⁷⁶

Calculations for the group 1 HIV protease calculations were based on PDB structure 1HVR,⁷⁷ which was solved with a bound cyclic urea inhibitor, while group 2 calculations used 2I0D,⁷³ which was solved with ligand AD81 of this group. Except as otherwise noted, all crystallographic solvent molecules were omitted. For group 1a ligands, an explicit flap water was included with a flat-bottomed restraint on the water's oxygen atom, with $R = 0.5$ Å and $k = 1.0$ kcal/mol/Å.¹² The location of the oxygen was obtained by superimposing the protein parts of structures 1HVR and 2FDE⁷⁸ (HIV protease with GW0385) and then using the water coordinates from the superimposed 2FDE structure. On the basis of prior analyses, Asp 25 was treated as protonated on oxygen, and Asp 25' was treated as unprotonated for the group 1a compounds,⁷⁹ while Asp 25 and 125 were both protonated for the group 1b compounds.⁸⁰ The initial bound conformation of each HIVP ligand was generated by superimposing the common parts of its structures with the highest affinity compound in its group.

Calculations for the PDE 10a inhibitors were based on PDB structure 2OVV,⁷⁴ which was solved with inhibitor **21** and is very similar to structure 2OVY,⁷⁴ which was solved with inhibitor **29**. Examination of these structures revealed that water 85 in 2OVV and 2OVY appears to play a particularly important bridging role, as it is the only water within 3 Å of both the ligand and the protein. Moreover, initial conformational searches with no crystallographic waters generated conformations in which the ligand wandered from its crystallographic location, whereas including water 85 with a flat-bottomed restraint ($R = 0.5$ Å, $k = 1.0$ kcal/mol/Å¹²) centered on its crystallographic site caused the ligands to remain close to their crystallographic locations. This water was therefore included in all subsequent calculations, with the same restraining potential. Initial conformations of the bound PDE 10a ligands were generated by superimposing them on the crystal conformation of the tightest-binding ligand, **29** in 2OVY. The bound Zn and Mg atoms were part of the real set and therefore treated as rigidly fixed. Both metals were treated as electrically neutral, as were

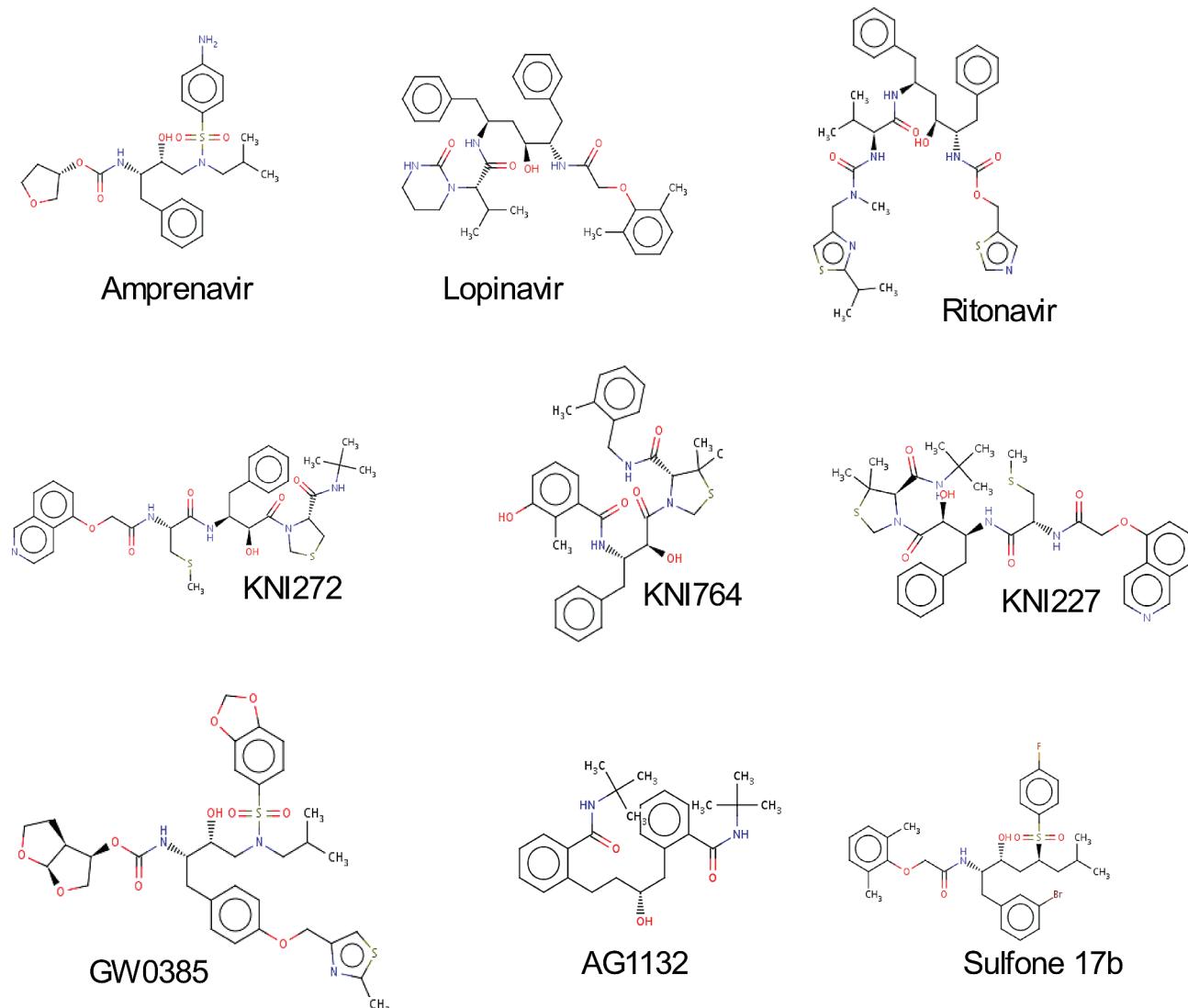


Figure 2. Group 1a HIV-1 protease inhibitors.

their first-shell histidine ligands. We do not know the actual protonation states of the first-shell ligands, but we observed that adding +2 charge to both metals (while keeping the first-shell histidines neutral) caused the conformational searches to yield grossly incorrect ligand poses. Fortunately, although the scaffolds of these compounds lie about 7–8 Å from the metals, the closest atoms of variable substituents (R in Figure 5) lie at least 12.7 Å away from the metals, on the basis of the available crystal structures, and the substituents are uncharged and mainly aromatic in nature. As a consequence, a detailed treatment of the metal cluster is not expected to be critical in this series.

For both HIVP and PDE 10a, the live set—the set of binding-site atoms treated as mobile—was defined as all atoms within 7 Å of any atom of any ligand. The real set—the set of protein atoms treated as rigid and thus supporting the live set—comprised all protein residues having any atom within 5 Å of any live-set atom. For the HIVP structures, the initial live and real sets based on these criteria were symmetrized, so that both protein monomers would have the same live and real sets. This was done by expanding, rather than contracting, the live and real sets. All other protein atoms were deleted, in order to reduce the size of the

nonbonded pairlist and thus speed the calculations. In order to diminish any initial stress in the starting conformations that might artificially drive the binding site away from its crystallographic starting conformation, the protein models were subjected to an initial relaxation step in which both the live and real sets were temporarily treated as live, except for the two metals, which were still held fixed. All protein residues having an atom within 5 Å of any live atom were temporarily treated as real, and the entire temporary live set was energy-minimized. Figures 6–8 highlight the live sets used in the present calculations. The program VMD⁸¹ was used for all 3D molecular graphics in this paper.

3. Results

3.1. Free Energy Calculations. *3.1.1. Group 1 HIVP Inhibitors.* The computed binding free energies for the group 1a and group 1b protease inhibitors are shifted relative to experimental results but correlate well (correlation coefficient $r \sim 0.8$), as shown by the scatter plots in Figure 9 and the linear regression parameters and mean errors in Table 1. Interestingly, the computed free energies of group 1b run lower than those of group 1a by about 10 kcal/mol (Figure

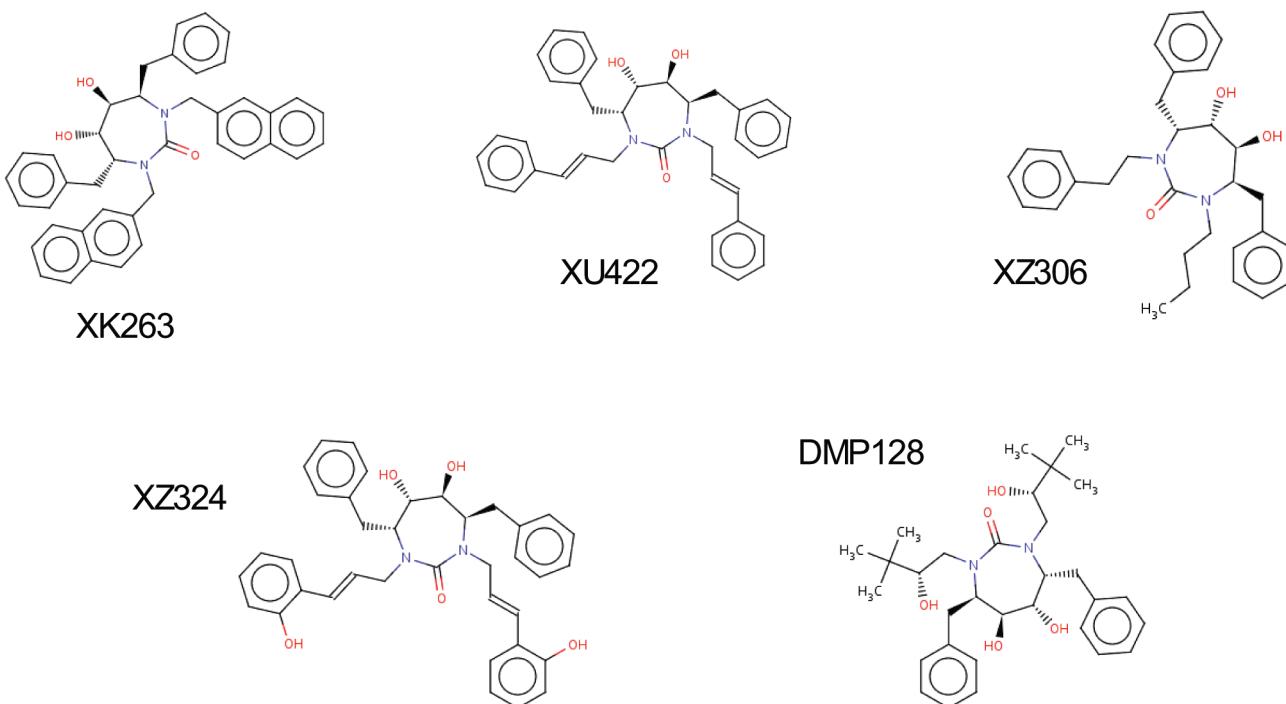


Figure 3. Group 1b HIV-1 protease inhibitors.

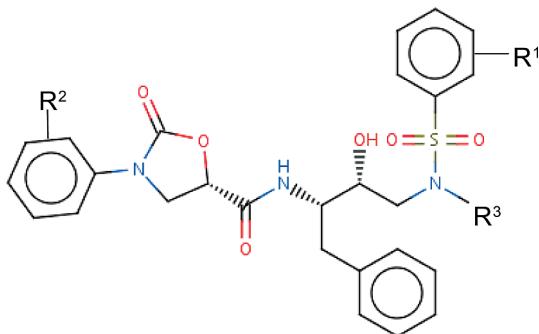


Figure 4. Scaffold of group 2 HIV-1 protease inhibitors.

9). We surmise that this offset stems primarily from the difference in the setup of the two protease structures. Thus, as described in the Methods section, the group 1a calculations include an explicit “flap” water molecule and treat one of the catalytic aspartyls as ionized, whereas the group 1b calculations omit the flap water and treat both aspartyls as neutral. Table 2 provides information on the root-mean-square deviation (RMSD) of the most stable computed conformations relative to crystallographic structures, where available. Results are presented for the ligand alone, and for the entire mobile region. The results are rather mixed, with ligand-only RMSDs ranging from 1.4 to 3.8. Examination of these structures indicates that the structural deviations are primarily associated with extended substituents, rather than with core parts of the ligand that interact with the central Asp groups and the flap Ile residue. Note that all group 1 calculations were based upon the same protein structure, 1HVR,⁷⁷ which was solved with bound XK263. It is also of interest to examine the conformations found in the calculations. Figures 10 and 11 give a sense for this range by overlaying representative conformations of the four most stable conformational clusters found for the bound states of

amprenavir (group 1a) and XK263 (group 1b). In both cases, the representative conformations may be viewed as minor modulations of a basic binding mode.

3.1.2. Group 2 HIVP Inhibitors. The group 2 inhibitors were all modeled on the basis of PDB structure 2I0D,⁷³ which was solved with bound AD81 from this series. The scatter plot of calculation versus experiment, Figure 12, shows nine nearly collinear points and one gross outlier, which corresponds to inhibitor KB19. Corresponding linear regression parameters, omitting KB19, are provided in Table 1. The mean deviation of calculation from experimental results for the group 2 inhibitors, −7.1 kcal/mol, is similar to that for the group 1a inhibitors, −5.1 kcal/mol, while the mean deviation of the group 1b inhibitors is considerably greater, at −16.5 kcal/mol. This similarity of the group 2 results to the group 1a results may stem from the fact that the group 1a and group 2 calculations both included an explicit flap water and treated the two catalytic aspartyls as having a net charge of −1. In contrast, the group 1b calculations did not include a flap water and treated the aspartyls as neutral. For those compounds in group 2 for which crystal structures are available, the RMSDs range from 1.9 to 2.4 for the ligand only, and from 2.0 to 2.8 for the entire mobile region. Compared with the RMSDs for group 1, these values are relatively moderate and uniform. This is not unexpected, given the comparative uniformity of the compound series. The variation of the computed structures across ligands is examined in Figure 13, which overlays the most stable bound conformation computed for each of the 10. The conformations are all quite similar, with most of the conformational differences localized in the variable substituents.

Examination of these bound conformations led to a potential explanation of why inhibitor KB19 is an outlier in the scatter plot (Figure 12). When the conformation of AD32,

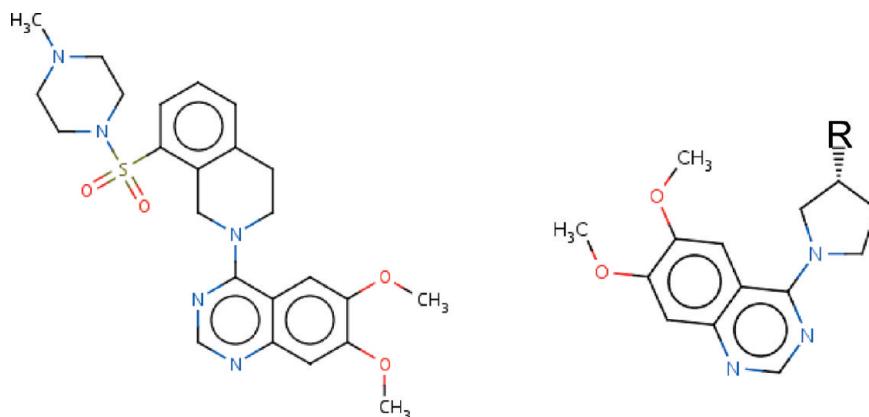


Figure 5. PDE 10a inhibitor **1** (left) and scaffold for inhibitors **11–29**.

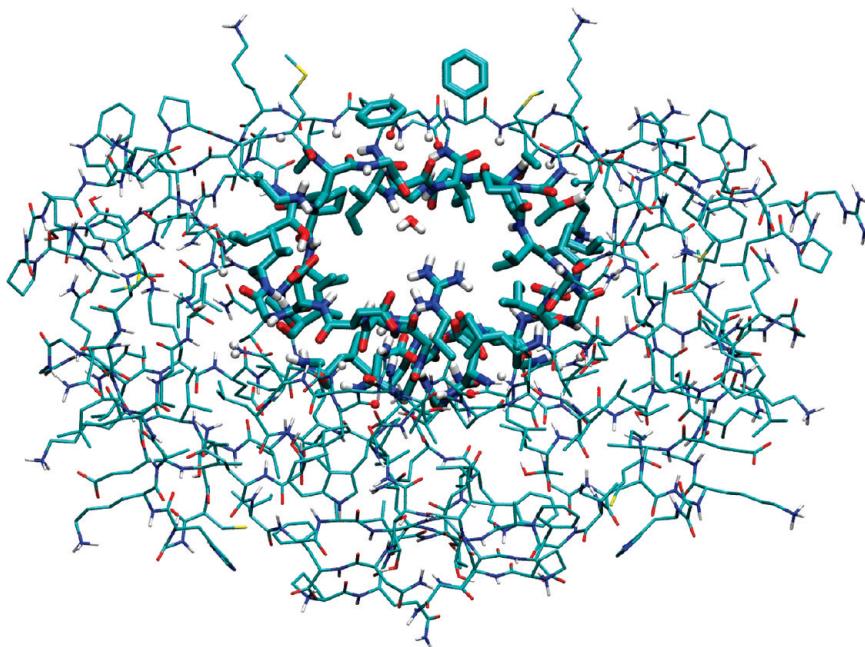


Figure 6. HIV-1 protease structures used for group 1 inhibitors, with mobile (“live”) set highlighted. (The inclusion in the live set of two apparently isolated phenyl rings results from the proximity of an inhibitor to one of the rings, combined with the symmetrization of the live set across the two protease monomers.) This and other 3D molecular graphics were generated with the program VMD.⁸¹

which also has a 4-acetyl substituent at the R² location, was compared, we noticed that this substituent lies in opposite orientations in the two compounds. That is, in the most stable conformations computed for these two compounds, the oxygen of each lay on the methyl carbon of the other. Because AD32 is not an outlier, we restarted the KB19 conformational search from a conformation modeled on the most stable conformation of AD32. This led to a conformation of the KB19 complex in which the 4-acetyl group is oriented as in AD32, and whose energy is about 8 kcal/mol more stable than any found before. (See second KB19 line (italic) in Table 2.) This lower energy brings KB19 largely into line with the other group 2 compounds (Figure 12) and lowers its RMSD somewhat (Table 2).

3.1.3. PDE 10a Inhibitors. Computed and measured binding free energies for the 20 congeneric PDE 10a inhibitors are compared in Table 3 and Figure 14, and the corresponding linear regression parameters and mean error are presented in Table 1. A significant correlation is observed,

and the correlation coefficient of 0.84 is similar to those obtained for the HIV protease inhibitors (above). However, the slope of the linear fit is about half as great as that obtained for the protease inhibitors. Figure 15 gives a sense of the conformational variation among the 20 bound ligands via an overlay of their most stable computed conformations, along with mobile residues in the PDE 10a binding site. The common scaffold occupies a uniform pose, while, not surprisingly, the variable substituent shows a wider range of positions. The side chains showing greatest conformational flexibility appear to be Met 251 and Met 252, which lie near the variable ligand substituent. The higher RMSD (Table 3) for compound **1**, relative to the other ligands with available crystal structures, **21** and **29**, may result from the fact that this compound is not based on the same chemical scaffold as any of the other compounds in the series (Figure 5) and that the protein structure used in the calculations was solved with one of the other compounds.

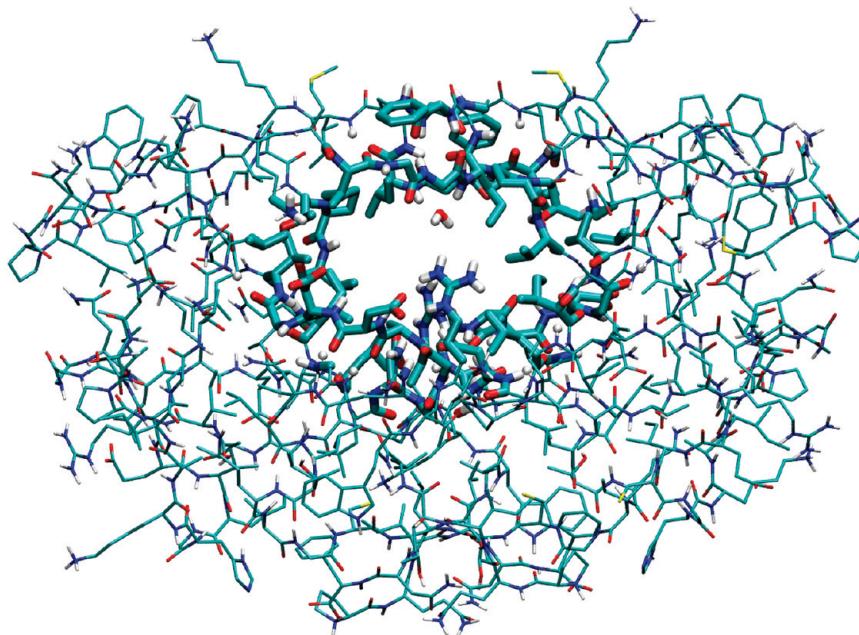


Figure 7. HIV-1 protease structure used for the group 2 inhibitors, with mobile (“live”) set highlighted.

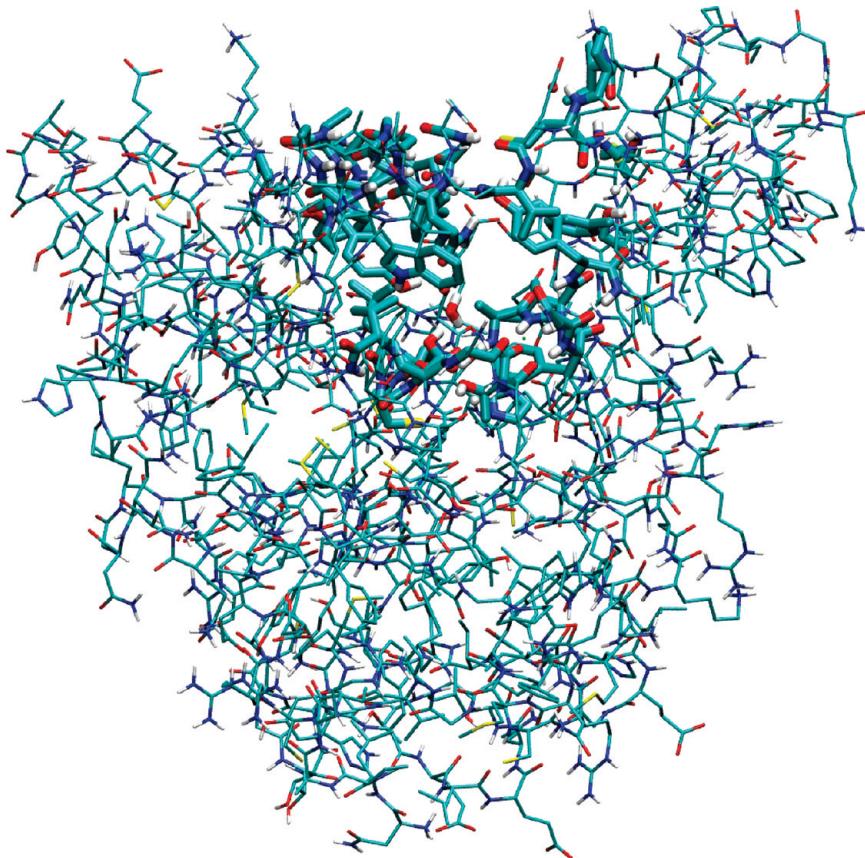


Figure 8. Phosphodiesterase 10a (PDE 10a) structure used for all the PDE 10a inhibitors, with mobile (“live”) set highlighted.

3.2. Entropy and Energy. The VM2 method yields not only overall binding free energies but also changes in mean energy components and in the configurational entropy. These breakdowns are provided in Tables 2 and 3 for the HIV protease and PDE 10a inhibitors, respectively. As in prior mining minima calculations on simpler host–guest systems, the net free energy changes are balances of considerably

larger energy and entropy contributions. In particular, the change in the Boltzmann averaged energy, $\Delta\langle E \rangle$, is considerably more negative than the binding free energy, ΔG° , while the configurational entropy makes a markedly positive contribution, $-T\Delta S^\circ$, and thus strongly opposes binding. The binding free energy can thus be viewed as a small difference between large numbers and is hence potentially subject to

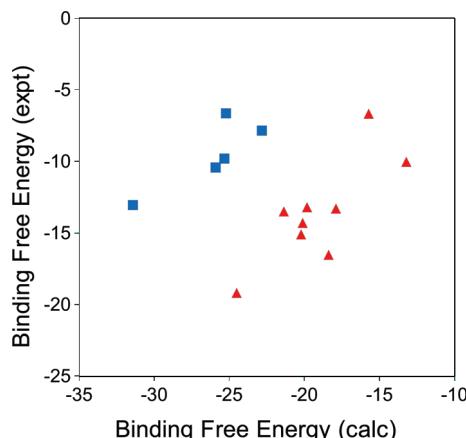


Figure 9. Experimental vs calculated binding free energies (kcal/mol) for group 1a (red triangles) and group 1b (blue squares) HIV-1 protease inhibitors.

Table 1. Linear Regressions and Mean Errors of Computed Binding Free Energies (ΔG°) versus Measured Binding Free Energies, Showing Correlation Coefficient (R), Slope and Y Intercept (kcal/mol) of Linear Regression Fits, and Mean Deviation (kcal/mol) of Calculation Relative to Experimental Results^a

	R	slope	Y intercept	mean error
HIVP group 1a	0.79	0.88	3.2	-5.5
HIVP group 1b	0.84	0.65	7.5	-16.6
HIVP group 2 ^b	0.98	0.56	-1.1	-7.1
PDE 10a	0.84	0.38	-8.3	5.2

^a In all cases, the experimental binding free energy is considered to be on the ordinate (y-axis). ^b Group 2 regressions omit the outlier, KB-19.

large uncertainties on this account. It is thus worth emphasizing that the mining minima algorithm does not generate the free energy by computing energy and entropy separately, and then adding them. Rather, it yields the free energy directly, and this can later be decomposed into entropy and energy. This integrated approach represents a potential advantage over methods which compute the two terms separately. It is also important to recognize that, although $\Delta\langle E \rangle$ omits the configurational entropy, it implicitly includes the change in solvent entropy via the implicit solvent model. As a consequence, the entropy values reported here should not be directly compared with experimental entropy changes of binding for these systems.

The computed entropic penalties range from 25 to 37 kcal/mol and average 30 kcal/mol. The energy changes are of greater magnitude and are negative in sign (average -44 kcal/mol), as indeed they must be if the computed binding free energies are to be negative, since $\Delta G^\circ = \Delta\langle E \rangle - T\Delta S^\circ$. In addition, the range of energy changes ($\Delta\langle E \rangle$), -63 to -27 kcal/mol, is about 3 times broader than the range of entropic contributions. Not surprisingly, both the entropy and energy contributions for these protein–ligand systems are considerably larger, by factors of ~2.8, than those computed previously for smaller host–guest systems,^{57,59,61,62} where $-T\Delta S^\circ$ and $\Delta\langle E \rangle$ averaged about 11 and -16 kcal/mol.

The importance of the configurational entropy to the correlations between calculation and experimentation re-

ported above may be examined by comparing the measured binding free energies with the computed values of $\Delta\langle E \rangle$ instead of ΔG° . Figures 16 and 17 provide such comparisons in graphical format for the HIV protease inhibitors, and Figure 18 provides the comparison for the PDE 10a inhibitors. The corresponding linear regression parameters and mean errors are provided in Table 4 (first 4 rows of data). (The group 2 regression again omits the outlier KB19.) Overall, the new correlation coefficients tend to be somewhat lower than those found for the full computed free energy, with the mean R lower by 15%, especially for PDE 10a. Even more striking are the large reductions in the regression slopes by a mean factor of 0.59, and the markedly more negative mean errors by about -30 kcal/mol. These systematic changes highlight the fact that the computed entropies strongly oppose binding and that the entropic penalty correlates with the binding energy, as further analyzed later in this section.

The values of $\Delta\langle E \rangle$ considered in the prior paragraph are Boltzmann-weighted averages, where the Boltzmann weighting is based upon the computed free energy of each energy well. Computing such average energies thus requires the full mining minima algorithm. It is also of interest to examine the even simpler energy-based approach of simply subtracting the global energy minima of the free protein and ligands from those of the protein–ligand complexes. The resulting energy changes, ΔE_{\min} , are not averages, and they entirely exclude any consideration of configurational entropy. However, they do require a conformational search and do include the PB/SA correction to the GB solvation energy (see the Methods section). These energy differences are compared with experimental binding free energies in the last four rows of Table 4. The results are much the same as for $\Delta\langle E \rangle$.

Prior mining minima calculations for host–guest systems yielded a striking correlation between configurational entropy and energy.^{57,59,61} It is thus of interest to put the energy and entropy results for the present protein–ligand calculations into the context of these prior host–guest results. As shown in Figure 19, the protein ligand results may be viewed as broadly extending the approximately linear entropy–energy relationship found for the host–guest systems. This observation is consistent with the observation that linear regressions of measured binding free energy versus computed binding energy yield lower slopes and more positive y intercepts than linear regressions of measured binding free energy versus computed binding free energy (Tables 2 and 4). This is because including the entropic contribution reduces the range of the computed values and makes the computed results more positive (opposing binding).

A more careful look at Figure 19 indicates that the PDE 10a data follow the host–guest trend particularly closely. The combined host–guest and PDE 10a data yield the following regression fit: $-T\Delta S^\circ = -0.88\Delta E - 1.56$. Since $\Delta G^\circ = \Delta E - T\Delta S^\circ$, this regression corresponds to the energy-free energy relationship $\Delta G^\circ = 0.12\Delta E - 1.56$, which indicates that most of the energetic driving force for binding is predicted to be canceled by a proportionate entropic penalty. The entropic penalties for the HIVP systems fall below the trend established by the host–guest and PDE 10a

Table 2. Detailed Energy (kcal/mol) Breakdowns of Computed HIVP-Inhibitor Binding Free Energies, $\Delta G^\circ(\text{calc})$, along with Experimental Binding Free Energies, $\Delta G^\circ(\text{expt})$; Change in Mean Energy Associated with Force-Field Bond-Stretch, Angle-Bend, and Dihedral Terms (Valence); Change in Mean Force-Field Coulombic Energy (Coulomb); Change in Mean Poisson–Boltzmann Solvation Energy (PB); Change in Mean Nonpolar Surface Energy (NP); Change in Mean Force-Field Lennard-Jones Energy (VDW); Change in Mean Total Energy (ΔE , the Sum of the Prior Five Terms); and Change in Configurational Entropy Contribution to the Free Energy ($-\Delta S^\circ$)^a

inhibitor	Boltzmann-averaged energy changes									
	$\Delta G^\circ(\text{expt})$	$\Delta G^\circ(\text{calc})$	valence	Coulomb	PB	NP	VDW	ΔE	$-\Delta S^\circ$	RMSD (Å)
group 1a										
Amprenavir	-13.2 ^b	-19.82	-0.15	-33.28	61.00	-6.60	-69.40	-48.43	28.61	1.4/3.7 ^k
Lopinavir	-15.1 ^b	-20.22	0.87	-49.32	88.00	-7.73	-88.33	-56.51	36.30	1.4/1.1 ^l
Ritonavir	-13.5 ^c	-21.37	14.26	-78.40	101.30	-7.61	-87.99	-58.45	37.08	3.8/1.6 ^m
KNI272	-13.3 ^b	-17.90	9.07	-53.65	90.00	-7.38	-90.88	-52.85	34.95	3.3/1.6 ⁿ
KNI764	-14.3 ^b	-20.12	1.72	15.08	29.62	-6.80	-85.03	-45.41	25.29	2.5/2.3 ^o
KNI227	-16.5 ^d	-18.39	11.58	-55.50	91.13	-7.54	-91.80	-52.13	33.74	
GW0385	-19.2 ^e	-24.51	8.23	-5.45	38.77	-7.24	-88.13	-53.83	29.31	2.2/1.3 ^p
AG1132	-6.7 ^f	-15.72	12.12	-61.17	71.65	-6.11	-60.81	-44.32	28.60	
Sulfone 17b	-10.0 ^g	-13.21	6.60	-37.88	81.15	-7.39	-84.94	-42.46	29.25	
group 1b										
XK263	-13.1 ^h	-31.43	-10.03	-35.08	46.10	-7.44	-55.99	-62.43	31.01	1.5/1.3 ^q
XU422	-7.9 ⁱ	-22.84	-10.73	3.80	22.84	-7.54	-60.85	-52.49	29.65	
XZ306	-9.8 ^j	-25.33	-7.22	-28.51	36.26	-6.65	-47.49	-53.61	28.28	
XZ324	-10.4 ^j	-25.92	-9.15	-51.16	56.29	-7.60	-49.51	-61.12	35.21	
DMP128	-6.6 ^j	-25.23	-4.40	24.82	7.64	-6.64	-74.31	-52.88	27.66	-
group 2 ⁱ										
AD23 (29a)	-11.6	-19.88	2.92	-69.42	86.32	-6.98	-67.43	-54.59	34.71	
AD32 (37f)	-10.4	-16.74	-0.34	-71.59	90.86	-7.61	-62.33	-51.00	34.25	
AD67 (27b)	-14.0	-22.18	4.57	-74.23	94.32	-7.47	-74.40	-57.22	35.04	
AD74 (25c)	-13.3	-22.64	2.37	-85.13	94.82	-7.02	-62.91	-57.85	35.22	
AD81 (21e)	-16.7	-27.51	7.52	-75.63	90.62	-7.49	-76.81	-61.78	34.27	2.1/2.0 ^r
KB02 (21a)	-13.8	-22.27	0.85	-69.60	84.38	-7.06	-65.50	-56.92	34.65	
KB19 (21f)	-15.7	-12.99	-1.03	-62.36	81.58	-7.28	-60.32	-49.41	36.41	2.4/2.2 ^s
KB19 (21f)	-15.7	-20.83	-2.18	-57.59	73.80	-7.30	-60.50	-53.79	32.95	2.1/2.0
KB56 (36c)	-12.7	-19.63	-1.26	-74.88	102.96	-7.33	-70.37	-50.88	31.25	
KB60 (26d)	-14.9	-24.69	4.98	-75.40	91.24	-7.45	-74.57	-61.20	36.50	1.9/2.8 ^t
KK98 (37b)	-9.3	-14.38	3.10	-63.73	88.10	-7.49	-70.78	-50.80	36.42	

^a RMSD: root-mean-square deviation of most stable computed conformation of the bound complex for ligand alone/all mobile atoms (Angstroms). Experimental free energies sourced as follows: ^b Ref 96. ^c Ref 97. ^d Ref 98. ^e Ref 99. ^f Ref 100. ^g Ref 101. ^h Refs 77, 77, 102, and 103. ⁱ BindingDB Entry 285. ^j Ref 73. Crystal structures as follows: ^k 1HPV.¹⁰⁴ ^l 1MUI.¹⁰⁵ ^m 1HXW.¹⁰⁶ ⁿ 1HPX.¹⁰⁷ ^o 1KZK.¹⁰⁸ ^p 2FDE.⁷⁸ ^q 1HVR.⁷⁷ ^r 2I0D.⁷³ ^s 2I0A.⁷³ ^t 3GI4.¹⁰⁹

systems, indicating that a smaller proportion of binding energy is canceled by entropy losses. Although it is tempting to interpret this result as implying that the HIVP systems tend to overcome entropic compensation, it might equally well indicate that keeping much of the protein rigid, as done in these calculations, has led to a disproportionate underestimate of entropic losses for HIVP, relative to PDE 10a. The HIVP points in the entropy-energy scatter plot also show considerably greater scatter than the PDE or host–guest results; the reasons for this are unclear.

3.3. Mode Scanning and Poisson–Boltzmann Corrections. The PB/SA solvation energy correction and the mode-scanning correction to the harmonic approximation in free energy in the energy well (see the Methods section) make the calculations somewhat more complicated and time-consuming. It is therefore of interest to examine the extent to which they contribute to accuracy. Accordingly, Table 5 presents linear regression analyses of measured binding free energies vs binding free energies computed without the mode scanning and PB/SA corrections, respectively. Omitting mode scanning has minimal effect on the results (Table 5, top), but omitting the PB/SA correction seriously damages the correlations (Table 5, bottom). This is clearly due to the

PB, rather than the SA part, given the comparatively small and uniform values of the latter (Tables 2, 3).

3.4. Convergence and Timings. The progress of the computed free energies of the various protein–ligand complexes as a function of the number of conformational search cycles (see the Methods section) is displayed in Figures 20–22 for the group 1 and group 2 HIV protease inhibitors and PDE 10a inhibitors, respectively. The graphs all trend downward because accounting for additional conformations in the mining minima method can only lower the overall free energy. Thus, the sudden downward jumps in the graphs result from the discovery of new conformations of markedly lower free energy as the search proceeds. Note that the absolute free energies displayed in these graphs are not physically interpretable on their own, because the free energies of the free proteins and their respective ligands have not been subtracted out.

The free energy calculations reported here required an average of about 4.5 h per search cycle on a 2.6 GHz dual-core AMD Opteron 2218 CPU with 4 GB of RAM. Roughly 80% of this time is spent on conformational search and 20% on calculating the free energies of individual energy wells. The CPU time for computing the

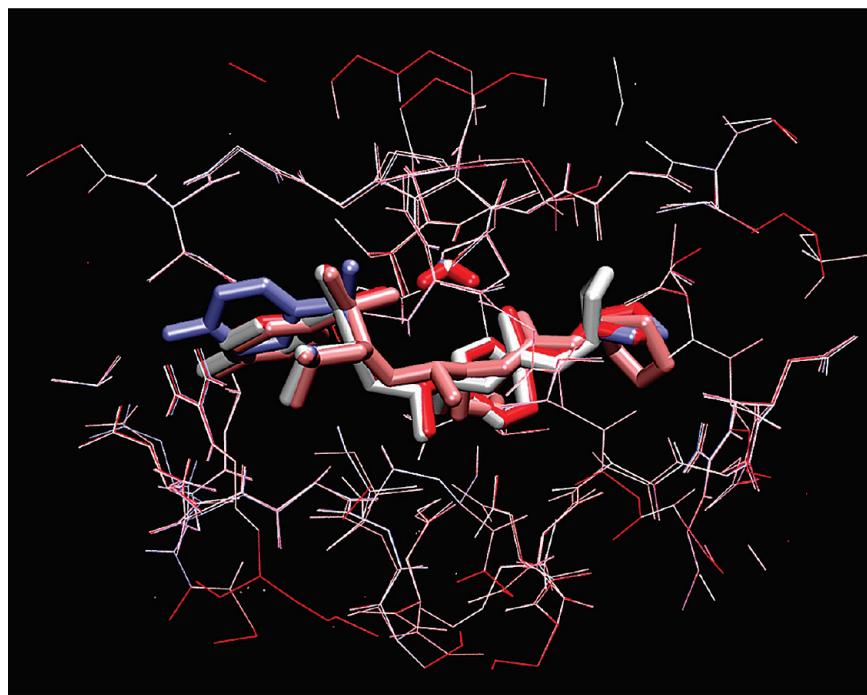


Figure 10. Lowest-energy conformers from each of the four most stable conformational clusters found for the complex of amprenavir with HIV-1 protease. Color indicates computed stability in the order red > pink > white > blue, from highest to lowest stability. Ligands and flap water shown as tubes, protein as lines.

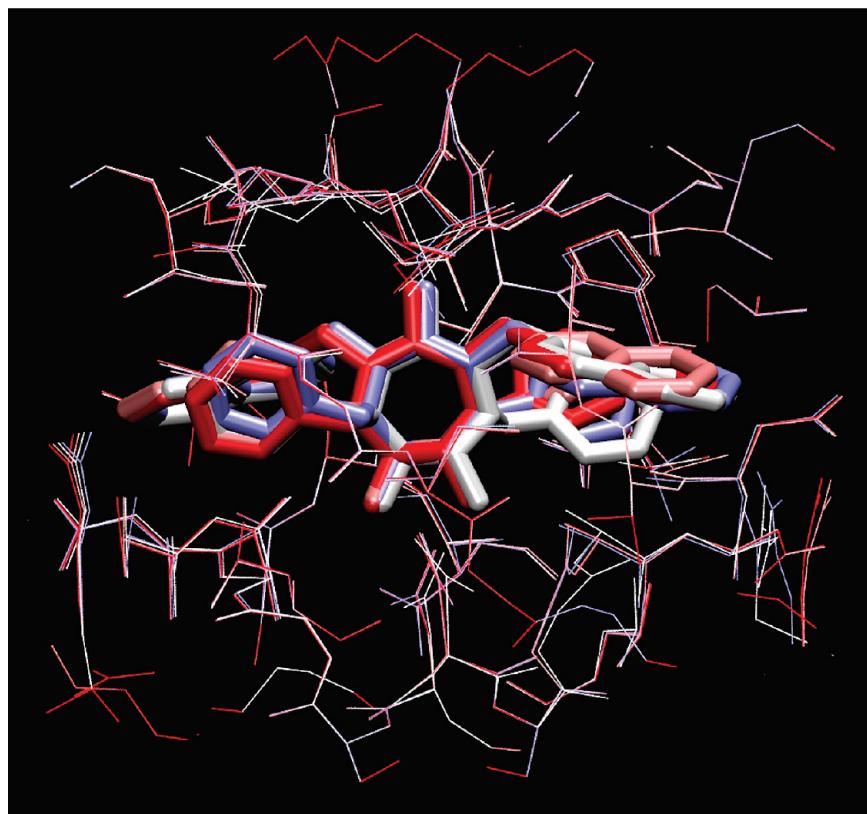


Figure 11. Lowest-energy conformers from each of the four most stable conformational clusters found for the complex of XK263 with HIV-1 protease. See caption of prior figure for details.

free energies of the individual energy wells partitions approximately as follows: 20% harmonic free energy approximation, 20% mode scanning, and 60% PB/SA solvation correction.

4. Discussion

These first applications of the mining minima approach to protein–ligand binding yield encouraging correlations be-

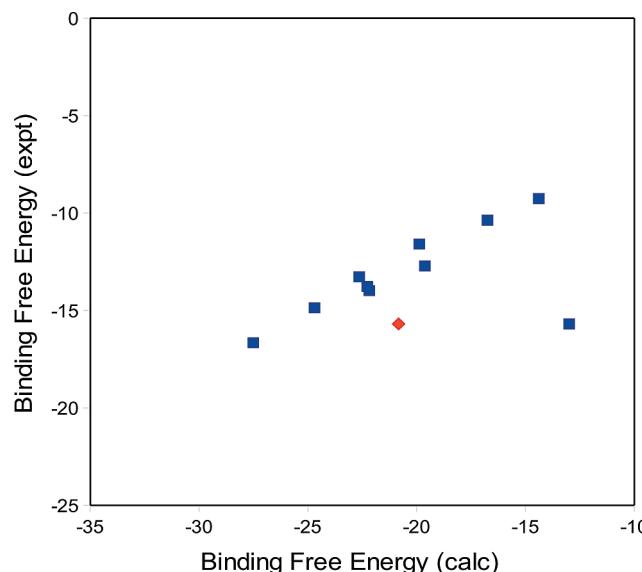


Figure 12. Measured vs computed binding free energies (kcal/mol) for group 2 HIV-1 protease inhibitors. Blue squares: initial calculations for all inhibitors. Red diamond: KB19 after recalculating the free energy of its complex based on the most stable conformation of AD32. (See text.)

tween calculated and measured binding free energies for inhibitors of two enzymes, HIV-1 protease and PDE 10a. There are several features of the present approach that may help it achieve these correlations. First, it is based upon a physics-based energy model comprising both an empirical force field and an implicit solvent model. Second, it accounts for the conformational flexibility of not only the ligand but also the side-chain and backbone atoms in the protein binding site. Third, the free energy of the ligand is computed and subtracted from the free energy of the complex, so that the computed binding free energies can account for the degree of ligand preorganization. Finally, the method accounts for changes in configurational entropy, not just energy.

The significance of the entropic contribution was tested by artificially omitting it. In the present calculations, neglecting the computed change in configurational entropy tends to weaken the correlation with experimental results, although this is not seen uniformly. In particular, omitting entropy for the group 1b HIVP inhibitors if anything slightly strengthens the correlation with experimental results. It is

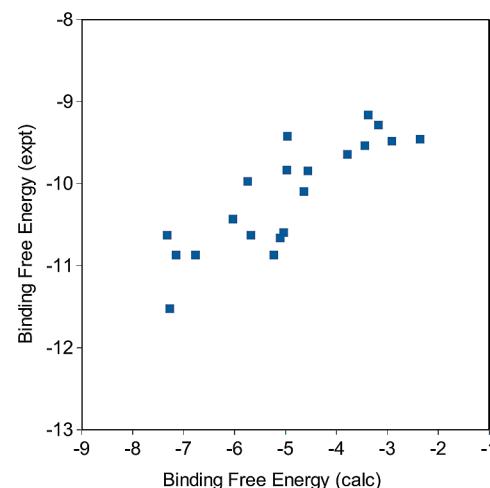


Figure 14. Experimental vs calculated binding free energies (kcal/mol) for 20 PDE 10a inhibitors.

interesting to speculate that this might be related to the greater rigidity and structural uniformity of these compounds. It is also of scientific interest that the calculations yield correlations between the change in configurational entropy on binding and the change in energy, very much like those seen in prior mining minima calculations for host–guest systems, and highly analogous to the entropy–enthalpy compensation seen in many experimental studies.^{82,83} This observation is consistent with the tendency of purely energetic physics-based models to overestimate the scale of protein–ligand binding affinities⁵⁹ since such models omit a systematically varying free energy penalty.

The limitations of this study also deserve comment. One is that we have simplified the challenge of achieving correlation with experimental results by focusing primarily on congeneric compound series and taking advantage of existing knowledge about each series, such as by including the flap water in two series of HIVP inhibitors. Also, the rather mixed structural RMSDs found for the group 1a HIVP inhibitors are a concern. On the other hand, highly accurate conformations may not be necessary in order to generate at least an approximate ranking of binding energies over the ~15 kcal/mol range spanned by this set. (The experimental affinity ranges are markedly smaller for group 2 and especially for the PDE 10a system.) It is also worth noting

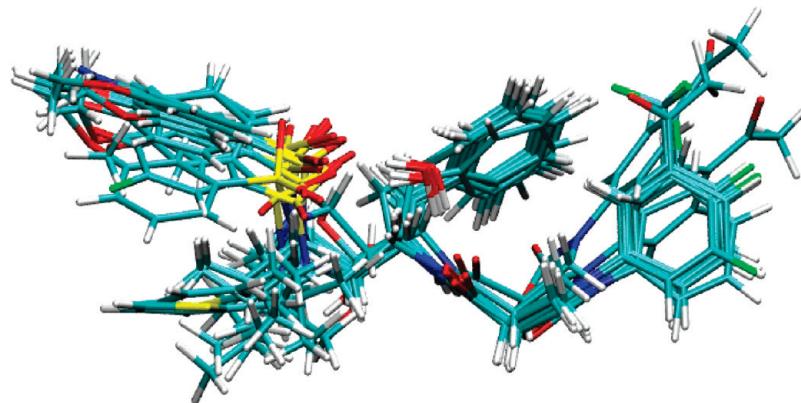
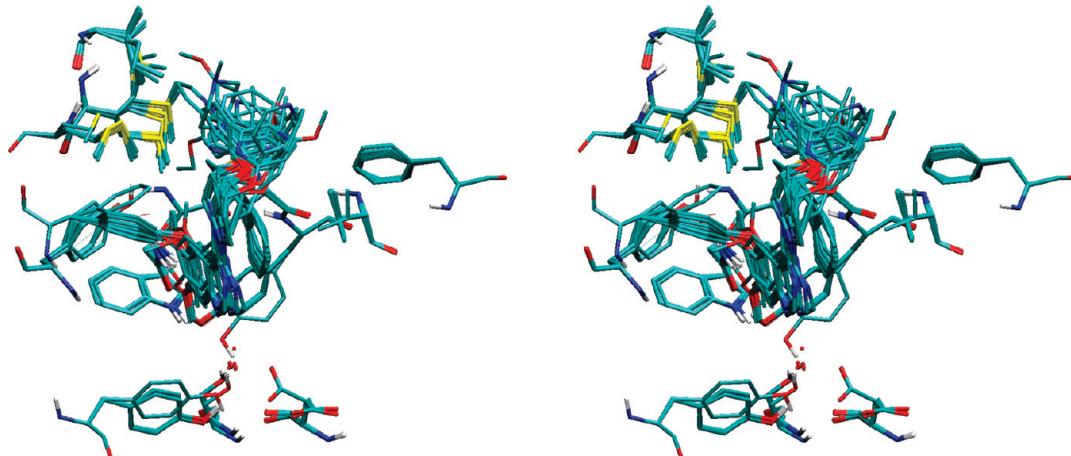


Figure 13. Most stable computed bound conformations of all 10 group 2 HIV-1 protease inhibitors, along with the flap water. Cyan: carbon. White: hydrogen. Red: oxygen. Blue: nitrogen. Green: fluorine. Yellow: sulfur.

Table 3. Detailed Energy Breakdowns for PDE 10a Inhibitors (see Table 2 for details)

inhibitor	Boltzmann-averaged energy changes									
	$\Delta G^\circ(\text{expt})^a$	$\Delta G^\circ(\text{calc})$	valence	Coulomb	PB	NP	VDW	ΔE	$-T\Delta S^\circ$	RMSD (Å)
1	-10.43	-6.03	4.52	-47.90	59.50	-5.03	-46.13	-35.04	29.01	3.0/1.6 ^b
11	-9.97	-5.74	-5.42	-31.69	41.45	-5.11	-33.23	-34.00	28.26	
12	-9.29	-3.17	0.50	-18.76	30.28	-4.78	-36.91	-29.68	26.50	
13	-9.42	-4.96	-2.73	-39.91	51.37	-5.31	-36.60	-33.17	28.21	
14	-9.64	-3.78	-2.71	-38.51	51.92	-4.84	-39.00	-33.15	29.36	
15	-9.84	-4.97	-2.92	-35.10	46.50	-4.71	-36.43	-32.66	27.69	
16	-9.85	-4.56	-2.34	-39.82	54.45	-5.02	-38.81	-31.54	26.98	
17	-10.10	-4.64	-6.17	-39.27	53.23	-4.60	-33.12	-29.93	25.29	
18	-9.48	-2.91	-6.70	-34.35	49.75	-4.54	-32.35	-28.19	25.28	
19	-9.46	-2.36	-2.70	-40.66	51.79	-4.35	-31.30	-27.22	24.86	
20	-9.54	-3.44	-4.32	-32.57	38.56	-4.34	-25.40	-28.08	24.63	
21	-10.87	-7.15	-3.68	-30.23	38.97	-4.92	-33.07	-32.94	25.79	1.2/1.2 ^c
22	-9.16	-3.38	-1.87	-28.31	35.93	-4.73	-32.33	-31.31	27.94	
23	-10.63	-5.68	-1.87	-42.53	53.22	-5.09	-38.74	-35.02	29.34	
24	-10.66	-5.10	-3.80	-39.06	50.73	-5.05	-38.12	-35.30	30.20	
25	-10.87	-6.77	-0.11	-36.38	42.57	-5.09	-34.23	-33.24	26.48	
26	-10.63	-7.32	-1.65	-33.18	40.46	-5.11	-34.49	-33.97	26.65	
27	-10.87	-5.23	-2.52	-39.64	51.38	-4.83	-35.81	-31.43	26.20	
28	-10.60	-5.03	-6.43	-33.56	49.85	-4.91	-35.60	-30.65	25.62	
29	-11.53	-7.27	-5.62	-39.45	54.14	-4.91	-37.82	-33.67	26.40	0.6/1.2 ^d

^a Ref 74. ^b 2O8H.⁷⁴ ^c 2OVV.⁷⁴ ^d 1OVY.⁷⁴**Figure 15.** Most stable bound conformations computed for the complexes of the PDE 10a inhibitors, showing the ligand surrounded by mobile residues in cross-eyed stereo.

that we have not simply redocked each ligand into a protein structure solved with that ligand but have instead used a single protein structure for all calculations within the ligand series. This approach provides a more challenging and real-world test of the methodology. Finally, this is a retrospective validation study, and a full evaluation of the method will require testing of prospective predictions. The results presented here suggest that it will be worth making such an effort.

The present calculations are time-consuming, yet, as with many other free energy methods, there are still concerns regarding convergence. The convergence problem is highlighted by the example of the HIVP inhibitor KB19, which appeared to be a gross outlier until we manually adjusted its conformation and restarted the conformational search, upon which its energy fell, and it came into line with the other group 2 inhibitors. It is essential to improve this aspect of the method so that one can be more confident of obtaining converged results. It is worth noting at the outset that the specific problem with KB19 could likely have been prevented

by adjustments to the run parameters of the existing conformational search algorithm, and, indeed, we are still defining “best practices” for the use of this software. At the same time, it will clearly be advantageous to speed up the calculations in order to accommodate more extensive conformational searches. One approach is to eliminate any unneeded, time-consuming steps and thus leave more time for the conformational search. Two possibilities were considered here: mode scanning and the PB/SA solvation correction. Omitting mode scanning, which provides a correction to the anharmonicity of the energy surface near the base of each energy well, made very little difference in the results, so it would seem reasonably safe to omit this part of the procedure. On the other hand, this step uses only about 2% of the total CPU time, and further testing might uncover cases where it is important. Omitting the PB/SA correction, on the other hand, seriously damaged the correlations of calculation with experimental results. Thus, although this step is computationally costly, it is indispensable. An alternative approach is to parallelize the calcula-

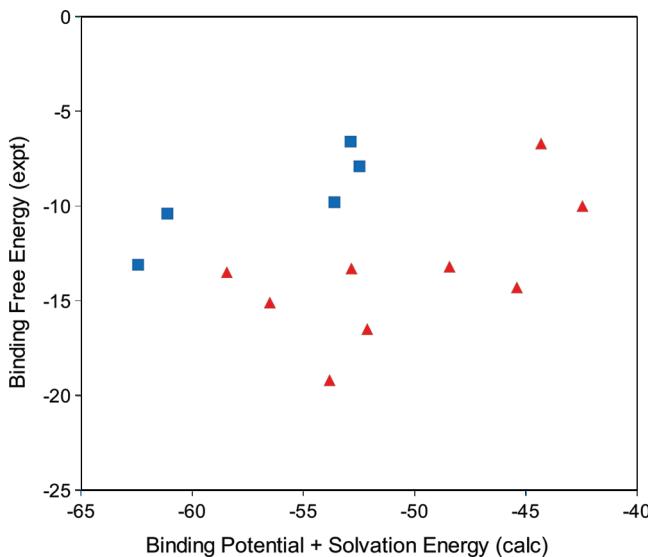


Figure 16. Measured binding free energy versus computed changes in potential plus solvation energy (omitting configurational entropy) for groups 1a (red triangles) and 1b (blue squares) HIV-1 protease inhibitors (kcal/mol).

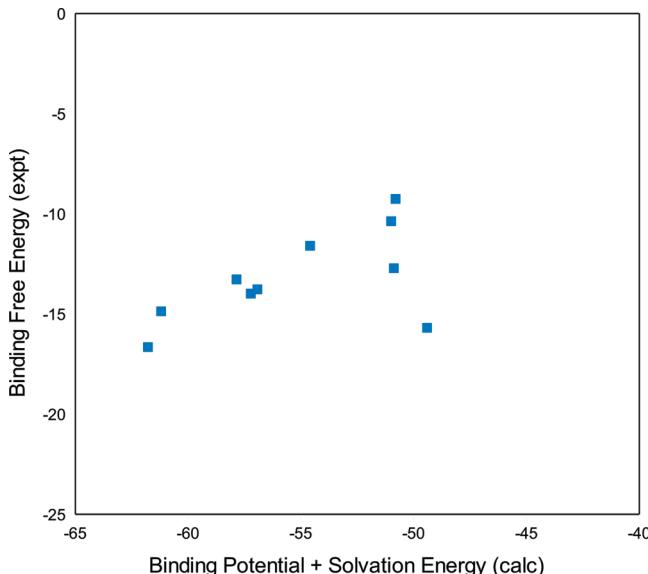


Figure 17. Measured binding free energy versus computed changes in potential plus solvation energy (omitting configurational entropy) for group 2 HIV-1 protease inhibitors (kcal/mol).

tions so that more extensive conformational searches can be carried out at low cost. Our preliminary efforts at fine-grained OpenMP parallelization and at porting key loops to graphical processor units (GPUs) indicate that substantial speedups can be achieved by these means. It will also be of interest to determine whether the search algorithm itself can be enhanced, such as by seeding it with candidate low-energy conformations generated with a simpler, faster algorithm.

It is of interest to compare the present mining minima method with the MMPBSA approach, because both are endpoint methods that use a force-field and an implicit solvent model to estimate binding free energies. One of the key differences is that mining minima simplifies the calculations by treating only the protein binding site as flexible. Treating

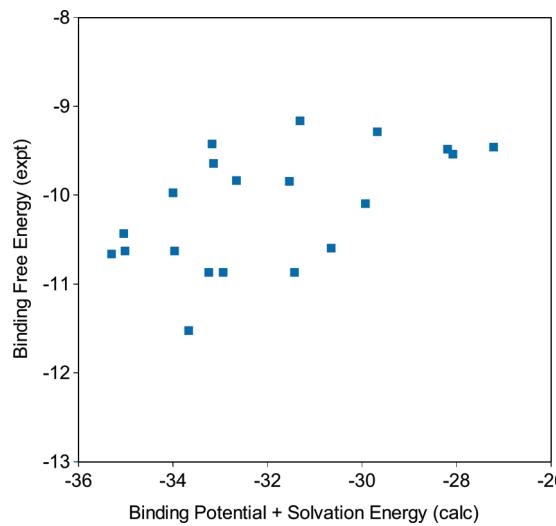


Figure 18. Measured binding free energy versus computed changes in potential plus solvation energy (omitting configurational entropy) for PDE 10a inhibitors (kcal/mol).

Table 4. Linear Regressions and Mean Errors of Computed Mean Binding Energies (Boltzmann-Averaged, $\Delta\langle E \rangle$ and Global Minimum, ΔE_{\min}) versus Measured Binding Free Energies, Showing Correlation Coefficient (R), Slope, and Y Intercept (kcal/mol) of Linear Regression Fits, and Mean Deviation (kcal/mol) of Calculation Relative to Experiment^a

	R	slope	Y intercept	mean error
$\Delta\langle E \rangle$				
HIVP group 1a	0.63	0.40	6.5	-37.0
HIVP group 1b	0.86	0.44	15.4	-47.0
HIVP group 2 ^b	0.89	0.47	13.5	-42.9
PDE 10a	0.56	0.15	-5.2	-21.9
ΔE_{\min}				
HIVP group 1a	0.67	0.42	7.3	-35.4
HIVP group 1b	0.89	0.48	16.3	-44.8
HIVP group 2 ^b	0.88	0.50	14.3	-40.7
PDE 10a	0.61	0.19	-4.3	-21.1

^a In all cases, the experimental binding free energy is considered to be on the ordinate (y axis). ^b Group 2 regressions omit the outlier, KB-19.

the entire protein as flexible, as is frequently done in MMPBSA calculations, in principle offers the possibility of accounting fully for protein-wide energy changes on binding. However, in practice, the large fluctuations of the energy of the entire protein makes it very difficult to converge the energy difference on binding. As a consequence, this contribution to the binding free energy is often completely discarded by using the single-trajectory MMPBSA approach.^{48,84} Focusing on the binding site allows mining minima to provide a detailed accounting of changes in the binding site energy, which is expected to be the most important part. Treating the protein as having a flexible binding site backed up by a shell of rigid atoms has the further benefit of allowing us to use conformational search algorithms that are much more aggressive than molecular dynamics, which is typically used in MMPBSA, without disrupting the overall protein structure. A second key difference is that MMPBSA calculations typically either neglect configurational entropy or else approximate it as an average vibrational entropy over

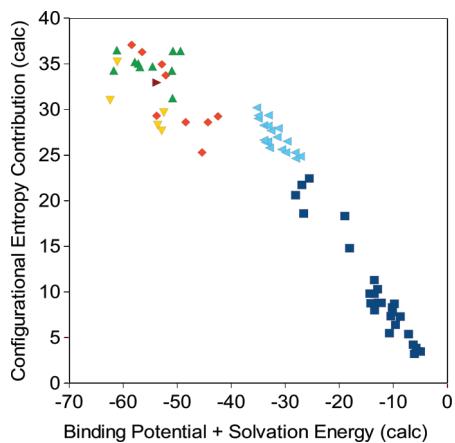


Figure 19. Computed configurational entropy contributions, $-T\Delta S^\circ$ vs energy contributions, ΔE , for the present protein–ligand calculations, along with prior host–guest results. Host–guest: blue squares. Group 1a HIVP inhibitors: red diamonds. Group 1b HIVP inhibitors: point-down yellow triangles. Group 2 HIVP inhibitors: point-up green triangles. PDE 10a: slanted, pale blue triangles. Recalculated KB19 (group 2) HIVP inhibitor: right-pointing brown triangle (kcal/mol).

Table 5. Linear Regressions of Experiment versus Computed Binding Free Energies with Neglect of Either the Mode Scanning Correction for Anharmonicity (Harmonic) or the PB/SA Solvation Correction (GB)

	R	slope	Y intercept
harmonic			
HIVP group 1a	0.81	0.89	2.5
HIVP group 1b	0.84	0.65	7.5
HIVP group 2*	0.97	0.56	-1.5
PDE 10a	0.83	0.37	-8.5
GB solvation			
HIVP group 1a	0.06	0.01	-12.5
HIVP group 1b	0.41	-0.08	-18.8
HIVP group 2*	0.52	0.12	-5.7
PDE 10a	0.21	0.029	-10.1

essentially randomly selected, energy-minimized molecular dynamics snapshots. We are not aware of any systematic effort to characterize or validate the MMPBSA approach. The mining minima approach, in contrast, is clearly tied to theory and has been numerically tested for small systems where reliable entropies can be obtained by brute-force methods.⁸⁶

Potential avenues for improving the accuracy of the method center on the representation of the protein and the choice of energy and solvent models. One issue is that the present calculations were simplified by omitting parts of the protein remote from the binding site. This approximation could become problematic for ionized ligands, and especially when comparing ionized ligands with neutral ones, because then long-ranged electrostatic interactions will become more important. It should be possible to address this omission efficiently and straightforwardly by using precomputed grids of potentials to account for the influence of remote parts of the protein on the mobile atoms of the ligand and protein in the binding site.⁸⁵ The force-field used here is a Dreiding-based model, with Vcharge partial charges for

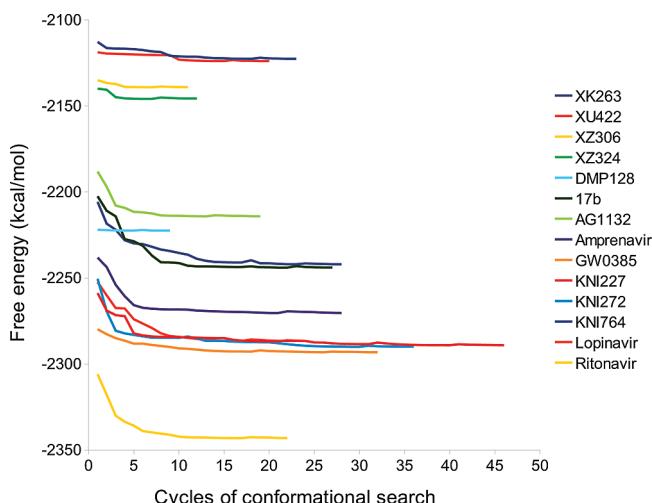


Figure 20. Convergence of computed free energies of protein–ligand complexes for group 1a and 1b HIV protease inhibitors, as a function of the number of search cycles (see Methods).

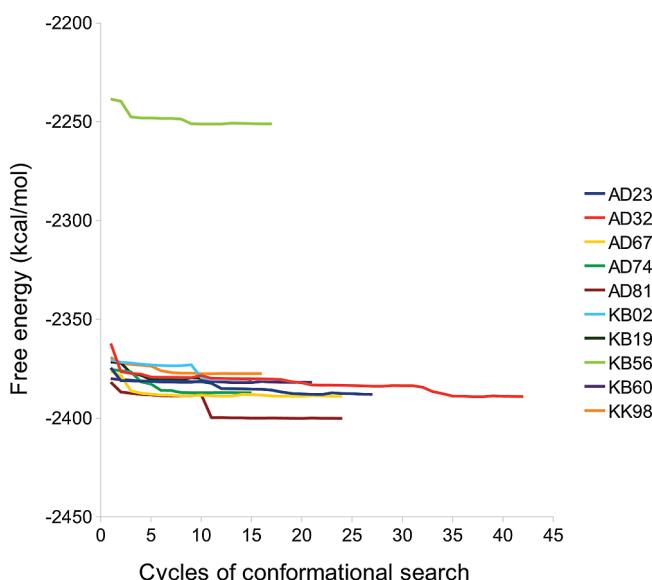


Figure 21. Same as prior figure, for group 2 HIV protease inhibitors. (The high energies of KB56 result from its cyclopropane ring, whose strain energy, evident here, cancels when the free energy of unbound KB56 is subtracted according to eq 1).

the inhibitors and CHARMM for the proteins. It is reasonable to ask whether more advanced energy models, such as ones that account for electronic polarizability,^{87–91} could increase reliability. More sophisticated energy models, potentially including quantum mechanical approaches,⁹² are likely to be important when metal atoms are closely involved in binding. It may be practical to include these as energy corrections as now done for the PB corrections, rather than having to pay the high computational cost of including them during the conformational search. Future calculations should also be able to take advantage of ongoing efforts aimed at assigning reliable parameters to a variety of drug-like ligands.^{93–95} Perhaps more critical is the treatment of the solvent. Here, we have effectively used a hybrid implicit-explicit model, since we have used knowledge of representa-

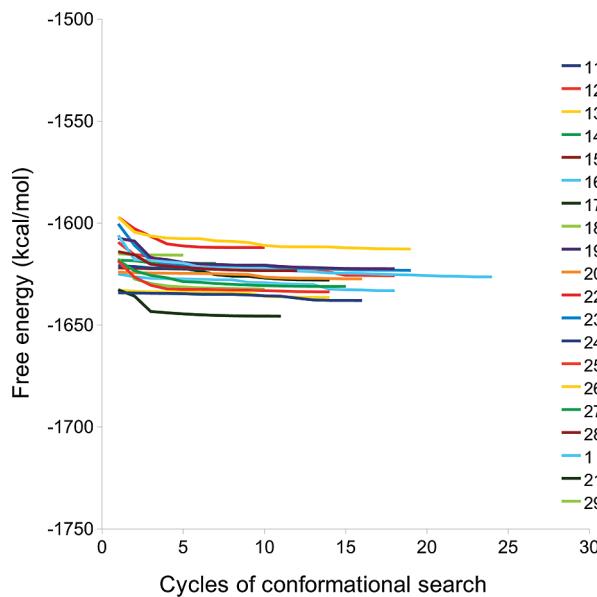


Figure 22. Same as prior two figures, for PDE 10a inhibitors. (The range of the vertical axis is matched to the prior figures to facilitate comparison.)

tive crystal structures to include key bridging waters explicitly. It would clearly be preferable to have an automatic means of accounting for the consequences of water's granular nature in the protein–ligand interface. One approach might be to routinely include one or a few explicit waters as additional "ligands" in the calculations. The present software implementation already allows this, and the added computational cost might remain manageable through parallelization, as discussed above. It should be noted, though, that such a procedure raises the theoretical issue of how to put the affinities of ligands computed with different numbers of explicit waters onto the same free energy scale. For example, part of the shift between the computed affinities of the present group 1a and group 1b HIVP inhibitors presumably results from the fact that group 1a includes one explicit water, while group 1b does not. We are currently working on this issue.

In summary, the mining minima approach has provided encouraging correlations between calculation with experimentation, and a number of avenues for further testing and improvement of the method have been identified. The approach appears to hold significant promise as a predictive method intermediate in complexity between fast, approximate docking and scoring methods and more time-consuming, rigorous free energy integration methods. It may thus be useful as a secondary computational filter for candidate ligands identified as promising by simpler, high-throughput docking methods, and as a tool to support the chemical optimization of lead compounds.

Acknowledgment. This work was made possible by grant R44GM075350 to M.J.P. from the National Institute of General Medical Sciences of the NIH. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Institute of General Medical Sciences.

Supporting Information Available: A zip file with the molecular topologies files for the drugs and drug-like

molecules studied in our paper, along with PDF document describing the parameters and units used in the topology files. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Chayen, N. *Struct. Genom., Part C* **2009**, *77*, 1–22.
- (2) Musa, T.; Ierger, T.; Sacchettini, J. *Struct. Genom., Part C* **2009**, *77*, 41–76.
- (3) Joachimiak, A. *Curr. Opin. Struct. Biol.* **2009**, *19*, 573–584.
- (4) Terwilliger, T.; Stuart, D.; Yokoyama, S. *Ann. Rev. Biophys.* **2009**, *38*, 371–383.
- (5) Shapiro, L.; Lima, C. *Structure* **1998**, *6*, 265–267.
- (6) Kim, S. *Curr. Opin. Struct. Biol.* **2000**, *10*, 380–383.
- (7) Congreve, M.; Marshall, F. *Br. J. Pharmacol.* **2010**, *159*, 986–996.
- (8) Rasmussen, S.; Choi, H.; Rosenbaum, D.; Kobilka, T.; Thian, F.; Edwards, P.; Burghammer, M.; Ratnala, V.; Sanishvili, R.; Fischetti, R.; Schertler, G.; Weis, W.; Kobilka, B. *Nature* **2007**, *450*, 383–U4.
- (9) Cherezov, V.; Rosenbaum, D.; Hanson, M.; Rasmussen, S.; Thian, F.; Kobilka, T.; Choi, H.; Kuhn, P.; Weis, W.; Kobilka, B.; Stevens, R. *Science* **2007**, *318*, 1258–1265.
- (10) Jaakola, V.; Griffith, M. T.; Hanson, M. A.; Cherezov, V.; Chien, E. Y. T.; Lane, J. R.; Ijzerman, A. P.; Stevens, R. C. *Science* **2008**, *322*, 1211–1217.
- (11) Palczewski, K.; Kumada, T.; Hori, T.; Behnke, C.; Motoshima, H.; Fox, B.; Le Trong, I.; Teller, D.; Okada, T.; Stenkamp, R.; Yamamoto, M.; Miyano, M. *Science* **2000**, *289*, 739–745.
- (12) Gilson, M.; Zhou, H. *Ann. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.
- (13) Villoutreix, B.; Eudes, R.; Miteva, M. *Comb. Chem. High Throughput Screen.* **2009**, *12*, 1000–1016.
- (14) Andricopulo, A.; Salum, L.; Abraham, D. *Curr. Top. Med. Chem.* **2009**, *9*, 771–790.
- (15) Prathipati, P.; Dixit, A.; Saxena, A. *Curr. Comput.-Aided Drug Des.* **2007**, *3*, 133–148.
- (16) Huang, N.; Jacobson, M. *Curr. Opin. Drug Discovery Dev.* **2007**, *10*, 325–331.
- (17) Lundstrom, K. *J. Cell. Mol. Med.* **2007**, *11*, 224–238.
- (18) Coupez, B.; Lewis, R. *Curr. Med. Chem.* **2006**, *13*, 2995–3003.
- (19) Warren, G.; Andrews, C.; Capelli, A.; Clarke, B.; LaLonde, J.; Lambert, M.; Lindvall, M.; Nevins, N.; Semus, S.; Senger, S.; Tedesco, G.; Wall, I.; Woolven, J.; Peishoff, C.; Head, M. *J. Med. Chem.* **2006**, *49*, 5912–5931.
- (20) Shoichet, B.; Bodian, D.; Kuntz, I. *J. Comput. Chem.* **1992**, *13*, 380–397.
- (21) Jones, G.; Willett, P.; Glen, R. *J. Mol. Biol.* **1995**, *245*, 43–53.
- (22) Oshiro, C.; Kuntz, I.; Dixon, J. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 113–130.
- (23) Jones, G.; Willett, P.; Glen, R.; Leach, A.; Taylor, R. *J. Mol. Biol.* **1997**, *267*, 727–748.

- (24) McMardin, C.; Bohacek, R. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 333–344.
- (25) Totrov, M.; Abagyan, R. *Prot. Struct. Funct. Genet.* **1997**, *215*–220.
- (26) Morris, G.; Goodsell, D.; Halliday, R.; Huey, R.; Hart, W.; Belew, R.; Olson, A. *J. Comput. Chem.* **1998**, *19*, 1639–1662.
- (27) Kramer, B.; Rarey, M.; Lengauer, T. *Prot. Struct. Funct. Genet.* **1999**, *37*, 228–241.
- (28) Gohlke, H.; Hendlich, M.; Klebe, G. *J. Mol. Biol.* **2000**, *295*, 337–356.
- (29) Raha, K.; Merz, K. *J. Am. Chem. Soc.* **2004**, *126*, 1020–1021.
- (30) Vasilyev, V.; Bliznyuk, A. *Theor. Chem. Acc.* **2004**, *112*, 313–317.
- (31) Grater, F.; Schwarzl, S.; Dejaegere, A.; Fischer, S.; Smith, J. *J. Phys. Chem. B* **2005**, *109*, 10474–10483.
- (32) Curioni, A.; Mordini, T.; Andreoni, W. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 773–784.
- (33) Raha, K.; Merz, K. *J. Med. Chem.* **2005**, *48*, 4558–4575.
- (34) Rajamani, R.; Good, A. *Curr. Opin. Drug Discovery Dev.* **2007**, *10*, 308–315.
- (35) Fong, P.; McNamara, J.; Hillier, I.; Bryce, R. *J. Chem. Inf. Model.* **2009**, *49*, 913–924.
- (36) Cho, A.; Chung, J.; Kim, M.; Park, K. *J. Chem. Phys.* **2009**, *131*.
- (37) Straatsma, T.; McCammon, J. *Annu. Rev. Phys. Chem.* **1992**, *43*, 407–435.
- (38) Tembe, B.; McCammon, J. *Comput. Chem.* **1984**, *8*, 281–283.
- (39) Wang, D.; Rizzo, R.; Tirado-Rives, J.; Jorgensen, W. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 2799–2802.
- (40) McCarrick, M.; Kollman, P. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 109–121.
- (41) Lau, F. T. K.; Karplus, M. *J. Mol. Biol.* **1994**, *236*, 1049–1066.
- (42) Dolenc, J.; Oostenbrink, C.; Koller, J.; van Gunsteren, W. *Nucleic Acids Res.* **2005**, *33*, 725–733.
- (43) Woo, H.; Roux, B. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6825–6830.
- (44) Mobley, D.; Chodera, J.; Dill, K. *J. Chem. Phys.* **2006**, *125*.
- (45) Fujitani, H.; Tanida, Y.; Ito, M.; Jayachandran, G.; Snow, C.; Shirts, M.; Sorin, E.; Pande, V. *J. Chem. Phys.* **2005**, *123*.
- (46) McDonald, J.; Brooks, C., III *J. Am. Chem. Soc.* **1991**, *113*, 2295–2301.
- (47) Reddy, M.; Viswanadhan, V.; Weinstein, J. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 10287–10291.
- (48) Brown, S.; Muchmore, S. *J. Chem. Inf. Model.* **2006**, *46*, 999–1005.
- (49) Gouda, H.; Kuntz, I.; Case, D.; Kollman, P. *Biopolymers* **2003**, *68*, 16–34.
- (50) Pearlman, D. *J. Med. Chem.* **2005**, *48*, 7796–7807.
- (51) Srinivasan, J.; Cheatham, T.; Cieplak, P.; Kollman, P.; Case, D. *J. Am. Chem. Soc.* **1998**, *120*, 9401–9409.
- (52) Aqvist, J.; Medina, C.; Samuelsson, J. *Protein Eng.* **1994**, *7*, 385–391.
- (53) Carlsson, J.; Ander, M.; Nervall, M.; Aqvist, J. *J. Phys. Chem. B* **2006**, *110*, 12034–12041.
- (54) JonesHertzog, D.; Jorgensen, W. *J. Med. Chem.* **1997**, *40*, 1539–1549.
- (55) Zhou, R.; Friesner, R.; Ghosh, A.; Rizzo, R.; Jorgensen, W.; Levy, R. *J. Phys. Chem. B* **2001**, *105*, 10388–10397.
- (56) Gilson, M.; Given, J.; Head, M. *Chem. Biol.* **1997**, *4*, 87–92.
- (57) Chang, C.; Gilson, M. K. *J. Am. Chem. Soc.* **2004**, *126*, 13156–13164.
- (58) Head, M.; Given, J.; Gilson, M. *J. Phys. Chem. A* **1997**, *101*, 1609–1618.
- (59) Chen, W.; Chang, C.; Gilson, M. K. *Biophys. J.* **2004**, *87*, 3035–3049.
- (60) Kolossvary, I. *J. Am. Chem. Soc.* **1997**, *119*, 10233–10234.
- (61) Chen, W.; Chang, C.; Gilson, M. K. *J. Am. Chem. Soc.* **2006**, *128*, 4675–4684.
- (62) Moghaddam, S.; Inoue, Y.; Gilson, M. *J. Am. Chem. Soc.* **2009**, *131*, 4012–4021.
- (63) Chang, C.; Gilson, M. K. *J. Comput. Chem.* **2003**, *24*, 1987–1998.
- (64) Kolossvary, I.; Guida, W. C. *J. Am. Chem. Soc.* **1996**, *118*, 5011–5019.
- (65) Keserû, G. M.; Kolossváry, I. *J. Am. Chem. Soc.* **2001**, *123*, 12708–12709.
- (66) Momany, F. A.; Rone, R. *J. Comput. Chem.* **1992**, *13*, 888–900.
- (67) Mayo, S.; Olafson, B.; Goddard, W. *J. Phys. Chem.* **1990**, *94*, 8897–8909.
- (68) Gilson, M.; Gilson, H.; Potter, M. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1982–1997.
- (69) Gilson, M.; Honig, B. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 5–20.
- (70) Qiu, D.; Shenkin, P.; Hollinger, F.; Still, W. *J. Phys. Chem. A* **1997**, *101*, 3005–3014.
- (71) Luo, R.; David, L.; Gilson, M. K. *J. Comput. Chem.* **2002**, *23*, 1244–1253.
- (72) Cai, Q.; Hsieh, M.; Wang, J.; Luo, R. *J. Chem. Theory Comput.* **2010**, *6*, 203–211.
- (73) Ali, A.; Reddy, G. S. K. K.; Cao, H.; Anjum, S. G.; Nalam, M. N. L.; Schiffer, C. A.; Rana, T. M. *J. Med. Chem.* **2006**, *49*, 7342–7356.
- (74) Chappie, T. A.; Humphrey, J. M.; Allen, M. P.; Estep, K. G.; Fox, C. B.; Lebel, L. A.; Liras, S.; Marr, E. S.; Menniti, F. S.; Pandit, J.; Schmidt, C. J.; Tu, M.; Williams, R. D.; Yang, F. V. *J. Med. Chem.* **2007**, *50*, 182–185.
- (75) Pan, Y.; Huang, N.; Cho, S.; MacKerell, A. D., Jr. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 267–272.
- (76) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L., III *J. Med. Chem.* **2004**, *47*, 3032–3047.
- (77) Lam, P.; Jadhav, P.; Eyermann, C.; Hodge, C.; Ru, Y.; Bacheler, L.; Meek, J.; Otto, M.; Rayner, M.; Wong, Y. *Science* **1994**, *263*, 380–384.
- (78) Miller, J. F.; Andrews, C. W.; Brieger, M.; Furfine, E. S.; Hale, M. R.; Hanlon, M. H.; Hazen, R. J.; Kaldor, I.;

- McLean, E. W.; Reynolds, D.; Sammond, D. M.; Spaltenstein, A.; Tung, R.; Turner, E. M.; Xu, R. X.; Sherrill, R. G. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 1788–1794.
- (79) Silva, A. M.; Cachau, R. E.; Sham, H. L.; Erickson, J. W. *J. Mol. Biol.* **1996**, *255*, 321–340.
- (80) Yamazaki, T.; Nicholson, L.; Torchia, D.; Wingfield, P.; Stahl, S.; Kaufman, J.; Eyermann, C.; Hodge, C.; Lam, P.; Ru, Y.; Jadhav, P.; Chang, C.; Weber, P. *J. Am. Chem. Soc.* **1994**, *116*, 10791–10792.
- (81) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graph.* **1996**, *14* (33–38), 27–28.
- (82) Lumry, R.; Rajender, S. *Biopolymers* **1970**, *9*, 1125–1227.
- (83) Sharp, K. *Protein Sci.* **2001**, *10*, 661–667.
- (84) Lee, M. S.; Olson, M. A. *Biophys. J.* **2006**, *90*, 864–877.
- (85) Given, J. A.; Gilson, M. K. *Proteins* **1998**, *33*, 475–495.
- (86) Chang, C.; Potter, M. J.; Gilson, M. K. *J. Chem. Phys. B* **2003**, *107*, 1048–1055.
- (87) Lopes, P.; Roux, B.; MacKerell, A. *Theor. Chem. Acc.* **2009**, *124*, 11–28.
- (88) Maple, J.; Cao, Y.; Damm, W.; Halgren, T.; Kaminski, G.; Zhang, L.; Friesner, R. *J. Chem. Theory Comput.* **2005**, *1*, 694–715.
- (89) Ren, P.; Ponder, J. *J. Comput. Chem.* **2002**, *23*, 1497–1506.
- (90) Cho, A.; Guallar, V.; Berne, B.; Friesner, R. *J. Comput. Chem.* **2005**, *26*, 915–931.
- (91) Kaminski, G.; Stern, H.; Berne, B.; Friesner, R.; Cao, Y.; Murphy, R.; Zhou, R.; Halgren, T. *J. Comput. Chem.* **2002**, *23*, 1515–1531.
- (92) Peters, M.; Raha, K.; Merz, K. *Curr. Opin. Drug Discovery Dev.* **2006**, *9*, 370–379.
- (93) Wang, J.; Wolf, R.; Caldwell, J.; Kollman, P.; Case, D. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
- (94) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; MacKerell, A. *J. Comput. Chem.* **2010**, *31*, 671–690.
- (95) Jakalian, A.; Jack, D.; Bayly, C. *J. Comput. Chem.* **2002**, *23*, 1623–1641.
- (96) Ohtaka, H.; Schon, A.; Freire, E. *Biochemistry* **2003**, *42*, 13659–13666.
- (97) Todd, M. J.; Luque, I.; Velazquez-Campoy, A.; Freire, E. *Biochemistry* **2000**, *39*, 11876–11883.
- (98) Kageyama, S.; Mimoto, T.; Murakawa, Y.; Nomizu, M.; Ford, H.; Shirasaka, T.; Gulnik, S.; Erickson, J.; Takada, K.; Hayashi, H. *Antimicrob. Agents Chemother.* **1993**, *37*, 810–817.
- (99) Hanlon, M. H.; Porter, D. J. T.; Furfine, E. S.; Spaltenstein, A.; Carter, H. L.; Danger, D.; Shu, A. Y. L.; Kaldor, I. W.; Miller, J. F.; Samano, V. A. *Biochemistry* **2004**, *43*, 14500–14507.
- (100) Melnick, M.; Reich, S. H.; Lewis, K. K.; Mitchell, L. J.; Nguyen, D.; Trippe, A. J.; Dawson, H.; Davies, J. F.; Appelt, K.; Wu, B.; Musick, L.; Gehlhaar, D. K.; Webber, S.; Shetty, B.; Kosa, M.; Kahil, D.; Andrade, D. *J. Med. Chem.* **1996**, *39*, 2795–2811.
- (101) Specker, E.; Bottcher, J.; Heine, A.; Sottriffer, C. A.; Lilie, H.; Schoop, A.; Muller, G.; Griebel, N.; Klebe, G. *J. Med. Chem.* **2005**, *48*, 6607–6619.
- (102) Hodge, C. N.; Aldrich, P. E.; Bacheler, L. T.; Chang, C.; Eyermann, C. J.; Garber, S.; Grubb, M.; Jackson, D. A.; Jadhav, P. K.; Korant, B. *Chem. Biol.* **1996**, *3*, 301–314.
- (103) Lam, P. Y. S.; Ru, Y.; Jadhav, P. K.; Aldrich, P. E.; DeLuca, G. V.; Eyermann, C. J.; Chang, C.; Emmett, G.; Holler, E. R.; Daneker, W. F.; Li, L.; Confalone, P. N.; McHugh, R. J.; Han, Q.; Li, R.; Markwalder, J. A.; Seitz, S. P.; Sharpe, T. R.; Bacheler, L. T.; Rayner, M. M.; Klabe, R. M.; Shum, L.; Winslow, D. L.; Kornhauser, D. M.; Jackson, D. A.; Erickson-Viitanen, S.; Hodge, C. N. *J. Med. Chem.* **1996**, *39*, 3514–3525.
- (104) Kim, E. E.; Baker, C. T.; Dwyer, M. D.; Murcko, M. A.; Rao, B. G.; Tung, R. D.; Navia, M. A. *J. Am. Chem. Soc.* **1995**, *117*, 1181–1182.
- (105) Stoll, V.; Qin, W.; Stewart, K. D.; Jakob, C.; Park, C.; Walter, K.; Simmer, R. L.; Helfrich, R.; Bussiere, D.; Kao, J.; Kempf, D.; Sham, H. L.; Norbeck, D. W. *Bioorg. Med. Chem.* **2002**, *10*, 2803–2806.
- (106) Kempf, D. J.; Marsh, K. C.; Denissen, J. F.; McDonald, E.; Vasavanonda, S.; Flentge, C. A.; Green, B. E.; Fino, L.; Park, C. H.; Kong, X. P. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 2484–2488.
- (107) Baldwin, E. T.; Bhat, T. N.; Gulnik, S.; Liu, B.; Topol, I. A.; Kiso, Y.; Mimoto, T.; Mitsuya, H.; Erickson, J. W. *Structure* **1995**, *3*, 581–590.
- (108) Reiling, K. K.; Endres, N. F.; Dauber, D. S.; Craik, C. S.; Stroud, R. M. *Biochemistry* **2002**, *41*, 4582–4594.
- (109) Nalam, M. N.; Ali, A.; Altman, M. D.; Reddy, G. S. K. K.; Chellappan, S.; Kairys, V.; Ozen, A.; Cao, H.; Gilson, M. K.; Tidor, B.; Rana, T. M.; Schiffer, C. A. *J. Virol.* **2010**, *02531–09*.

CT100245N