JCTC Journal of Chemical Theory and Computation

Linear Interaction Energy (LIE) Models for Ligand Binding in Implicit Solvent: Theory and Application to the Binding of NNRTIs to HIV-1 Reverse Transcriptase

Yang Su,[†] Emilio Gallicchio,*[,†] Kalyan Das,[‡] Eddy Arnold,[‡] and Ronald M. Levy*[,†]

BioMaPS Institute of Quantitative Biology, Department of Chemistry and Chemical Biology, and Center for Advanced Biotechnology and Medicine, Rutgers University, Piscataway, New Jersey 08854

Received August 8, 2006

Abstract: Expressions for Linear Interaction Energy (LIE) estimators for the binding of ligands to a protein receptor in implicit solvent are derived based on linear response theory and the cumulant expansion expression for the free energy. Using physical arguments, values of the LIE linear response proportionality coefficients are predicted for the explicit and implicit solvent electrostatic and van der Waals terms. Motivated by the fact that the receptor and solution media may respond differently to the introduction of the ligand, a novel form of the LIE regression equation is proposed to model independently the processes of insertion of the ligand in the receptor and in solution. We apply these models to the problem of estimating the binding free energy of two non-nucleoside classes of inhibitors of HIV-1 RT (HEPT and TIBO analogues). We develop novel regression models with greater predictive ability than more standard LIE formulations. The values of the regression coefficients generally conform to linear response predictions, and we use this fact to develop a LIE regression equation with only one adjustable parameter (excluding the intercept parameter) which is superior to the other models we tested and to previous results in terms of predictive accuracy for the HEPT and TIBO compounds individually. The new models indicate that, due to the different effects of induced steric strain of the receptor, an increase of ligand size alone opposes binding for ligands of the HEPT class, whereas it favors binding for ligands of the TIBO class.

1. Introduction

The binding free energy of a ligand to a receptor is given by the difference of the free energies of inserting the ligand in the receptor and in solution. In principle the free energy for each process can be calculated exactly for a given force field using the free energy perturbation (FEP) or thermodynamic integration (TI) methods. In practice, however, the complexities involved in setting up suitable mutation paths and the long simulation times needed to reach convergence have limited the applicability of FEP and TI methods to the investigation of the variations of the binding free energy for small ligand modifications in the final stages of lead optimization.[1] Linear Interaction Energy (LIE) models[2,3] offer attractive approximate alternatives to the full FEP methodology because they require only the computation of average interaction energies at the end points of the mutation.

LIE models can be described as empirical Quantitative Structure−Activity Relationships (QSAR) which employ physically motivated energetic estimators. As opposed to methods that predict the binding free energy on the basis of the structure of the ligand alone, LIE estimators also reflect properties of the ligand−receptor complex. LIE methods are expected to perform better than methods based on the ligand

* Corresponding author e-mail: emilio@biomaps.rutgers.edu (E.G.) and ronlevy@biomaps.rutgers.edu (R.M.L.).

[†] BioMaPS Institute of Quantitative Biology and Department of Chemistry and Chemical Biology.

[‡] Center for Advanced Biotechnology and Medicine and Department of Chemistry and Chemical Biology.

10.1021/ct600258e CCC: $37.00
Published on Web 11/10/2006

alone because they are more intimately related to the structure of the complex. One of the most popular LIE formulations employs the following regression expression for the binding free energy ΔF_b[3,4]

$$\Delta F_b = \alpha \Delta \bar{V}_{vdw} + \beta \Delta \bar{V}_{el} + \gamma \Delta \bar{A} + \delta \quad (1)$$

where $\Delta \bar{V}_{vdw}$, $\Delta \bar{V}_{el}$, and $\Delta \bar{A}$ are differences between quantities measured for the ligand complexed with the receptor and the ligand free in solution. $\bar{V}_{vdw}$ is the average van der Waals interaction energy between the ligand and its environment (the solvent or the receptor and the surrounding solvent). Similarly, $\bar{V}_{el}$ is the average electrostatic interaction energy between the ligand and its environment. $\bar{A}$ is the average solvent accessible surface area of the ligand. These quantities are typically obtained from Molecular Dynamics (MD) or Monte Carlo (MC) simulations started from known or modeled conformations of the ligand and the ligand−receptor complex. α, β, γ, and δ are empirical adjustable parameters whose values are obtained by fitting the model over a training set of ligands of known binding affinity. A trained LIE model can then be used to predict the binding free energy of ligands of unknown affinity provided that the LIE estimators for these ligands can be reliably calculated. Each LIE estimator in eq 1 reflects physical forces that affect binding. $\Delta \bar{V}_{vdw}$ and $\Delta \bar{V}_{el}$ measure the balance between the desolvation penalty caused by the loss of ligand−solvent interactions and the gain of ligand−receptor interactions which favor binding, whereas the surface area estimator measures the hydrophobic driving force toward complexation. These LIE estimators do not take into account explicitly thermodynamic forces related to the receptor that affect the binding affinity, such as the desolvation of receptor atoms and reorganization free energy of the receptor for accommodating the ligand. Nevertheless it can be shown (see below) that, under the assumption of linear response, these effects are, in fact, included in the model and are encoded in the values of the LIE regression coefficients.

LIE models have their origin in physical theories of solvation based on the linear response approximation to the free energy,[5−7] applied to the problem of binding free energy estimation.[2,8−11] The introduction of the ligand in either the solution or receptor environments can be regarded as a perturbation applied to the system. If the system responds perfectly linearly (as formally defined below) to the perturbation, the free energy of introducing the ligand can be shown to be exactly proportional to the interaction energy between the ligand and its environment. Given that the ligand constitutes a large perturbation to the system, it is unlikely that linear response applies to the entire processes of introducing the ligand in solution and in the receptor. The LIE regression equation (eq 1) assumes instead that linear response applies to the individual processes of introducing hydrophobic, van der Waals, and electrostatic interactions, albeit with different linear response proportionality coefficients. Even so, nonlinearities in practice limit the applicability of LIE models to within a related class of ligands. This is reflected in the fact that in practice LIE relationships are used for estimating *relative* binding free energies of similar ligands rather than absolute binding free energies. The accuracy of a LIE model therefore hinges on whether the perturbation corresponding to the mutation of each ligand into another is small enough so that linear response applies to the relative binding free energy. The effect of absolute binding free energies is collectively absorbed by the intercept parameter δ; deviations from linear response are expected to be reflected in the limited range of applicability of a LIE model.

In this paper we investigate a series of outstanding issues with regard to LIE models and their applications to ligand binding in structural biology. The first question is to what extent linear response applies to a given ligand set and the consequences of deviations from linear response in terms of the accuracy of the LIE model. Although in principle addressing this question requires comparing LIE predictions with relative free energies evaluated using the rigorous FEP and TI methods, we take a first step in this direction by comparing the values of the LIE adjustable parameters obtained by fitting training sets of ligand binding data with those expected based on linear response theory. The second question concerns the form of the LIE regression equation. Equation 1 assumes that the response of the solution and receptor environments, as measured by the LIE coefficients α, β, and γ, is the same. To our knowledge this assumption is ubiquitous in LIE applications. We develop and validate an alternative formulation in which the processes of insertion of the ligand in the solution and in the receptor are decoupled so that each is allowed to have different linear response proportionality coefficients. Finally, based on linear response techniques we analyze the appropriate form of the LIE estimators when the solvent is treated implicitly. We show that the form for the electrostatic LIE estimator we derive based on linear response theory agrees with the corresponding estimator recently proposed by Carlsson et al.[12] and differs from the more empirical expression proposed earlier by Zhou et al.[4] We develop, following the linear response formalism, electrostatic, van der Waals, and cavity implicit solvent estimators that best represent the corresponding LIE estimators in explicit solvent and show that these lead to improved accuracy.

We apply these ideas to ligand−protein complexes that have been studied previously using the LIE method: the binding of the HEPT and TIBO classes of Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTI) to the HIV-1 Reverse Transcriptase (RT) enzyme.[13] HIV-1 RT is essential for the life cycle of the virus. It converts the single stranded genomic RNA into double stranded DNA which is subsequently integrated into the host chromosome and passed on to all progeny cells.[14] Computer-aided structure-based drug discovery technologies have made a significant contribution to the development of medicinally active NNRTI anti-AIDS compounds.[15−17] Three of these compounds, dapivirine, etravirine, and rilpivirine, are currently in clinical trials. The goal is the design of NNRTIs of greater potency and resilience with respect to common drug-resistance mutations.[18] However the mode of binding of NNRTIs, which includes extensive receptor conformational reorganization and mainly nonspecific hydrophobic ligand−receptor contacts, does not offer obvious chemical modification leads for

binding free energy optimization. In this context, LIE models have been applied to the estimation of the binding affinities of NNRTI inhibitors with encouraging results.[19,20]

The renewed interest of our laboratories in the LIE modeling of HIV-1 RT inhibitors is motivated in part by recent progress in obtaining a crystal structure of HIV-1 RT in complex with an inhibitor of the N-acyl hydrazone class.[21] The crystal structure identifies a novel non-nucleoside binding site adjacent to but distinct from the NNRTI binding site. The NAH inhibitors targeting this site are expected to suffer from little or no cross-resistance from existing drug resistance mutations, thus providing new options for novel therapeutic strategies in the treatment of AIDS. This crystal structure provides the initial framework for the application of LIE methodologies for lead optimization and the development of a new class of inhibitors of HIV-1 RT. The work presented in this paper on NNRTIs provides the theoretical and computational basis for the application of LIE modeling in implicit solvent for binding free energy estimation of the new NAH class of HIV-1 RT inhibitors which is currently being investigated in our lab.

In the following section we review the theoretical foundations of the LIE method and discuss some of the approximations. We then derive a LIE formalism appropriate for situations where the solvent is modeled implicitly. We propose novel LIE regression equations which emerge naturally from the statistical mechanics derivation. We then apply these regression equations to analyze the binding of a series of 20 HEPT inhibitors and 37 TIBO inhibitors of HIV-1 RT. The models are validated using jack-knife prediction tests. The predictive ability of different forms and parametrizations of the LIE equations are compared. We compare features of the HEPT and TIBO binding modes and the corresponding LIE estimators. We conclude the paper with a discussion of the accuracy and physical interpretation of LIE models.

2. Theory and Methods

2.1. Linear Response Approximation. The idea of adopting a linear response approximation expression for binding free energy estimation was first stated by Lee et al.,[8] who suggested the use of a "two-point" Linear Response Approximation (LRA) estimation formula previously derived for electrostatic solvation;[6] see, for example, discussion in ref 10. This formula was later simplified by Åqvist et al.[2] who introduced the Linear Interaction Energy (LIE) model which, unlike the more accurate "two-point" LRA formula,[22] does not require the evaluation of estimators at more than one state of the ligand. The LIE method was used by Åqvist and collaborators to estimate relative binding affinities of endothiapepsin and HIV-protease inhibitors.[2,9] We review here some basic concepts related to the linear response approximation and derive a linear response expression—eq 12 below—for the LIE proportionality coefficients, which will be used in the following to interpret the values of the LIE fitting coefficients obtained from the analysis of experimental binding affinities.

The insertion of the ligand into either the receptor or the solvent can be thought of as turning on, by means of a charging parameter λ, the interactions of the ligand with the surrounding environment (the receptor and/or the solvent). It is convenient to think about these processes in stages. First the ligand cavity is formed, then the van der Waals interactions between the ligand and the environment are turned on, and finally ligand-environment electrostatic interactions are established. At each stage the λ-dependent potential energy of the system is

$$U(x;\lambda) = U_0(x) + \lambda V(x) \tag{2}$$

where V represents the ligand-environment interactions which are being added, and U_0, the reference potential energy, contains receptor−receptor, solvent−solvent, and ligand−intramolecular interactions as well as the ligand−environment interactions established in the previous stages. Starting from the expression of the configurational partition function

$$Z(\lambda) = e^{-\mathrm{F}(\lambda)/\mathrm{kT}} = \int dx\, e^{-\lambda \mathrm{V(x)/kT}} e^{-\mathrm{U_0(x)/kT}} \tag{3}$$

it is straightforward to show that the first and second derivatives of the free energy $F(\lambda) = -kT \ln Z(\lambda)$ with respect to λ correspond, respectively, to the first and second moments of the probability distribution of V:

$$\frac{\partial F}{\partial \lambda} = \langle V \rangle_\lambda \tag{4}$$

$$\frac{\partial^2 F}{\partial \lambda^2} = -\langle (\delta V)^2 \rangle_\lambda / kT \tag{5}$$

If the second moment is approximately constant along the thermodynamic path from $\lambda = 0$ to $\lambda = 1$, the third and higher order derivatives of F can be neglected, and it is possible to express $F(\lambda)$ as a Taylor series (known as the cumulant expansion of the free energy) centered at $\lambda = 0$ and truncated at the second order

$$F(\lambda) - F(0) = \langle V \rangle_0 \lambda - \frac{c}{2}\lambda^2 \tag{6}$$

where

$$c = \langle (\delta V)^2 \rangle / kT \tag{7}$$

is assumed constant for $0 \le \lambda \le 1$. It is of interest to note that when the fluctuations of the interaction potential V are Gaussian distributed this assumption is verified and eq 6 is exact.[6] According to eq 6 and under these assumptions, the free energy is quadratic with respect to the charging parameter. This is a manifestation of linear response behavior defined as when the solute-environment average interaction energy $\langle V \rangle_\lambda$ is linearly related to the charging parameter. Indeed, using eqs 4 and 5, and the same assumptions that have lead to eq 6, we obtain

$$\langle V \rangle_\lambda = \langle V \rangle_0 - c\lambda \tag{8}$$

which confirms that $\langle V \rangle_\lambda$ is linearly related to λ.

By evaluating eqs 6 and 8 at $\lambda = 1$ we obtain the free energy change and average interaction energy for adding solute-environment interactions under the assumption of linear response

$$\Delta F = F(1) - F(0) = \langle V \rangle_0 - \frac{c}{2} \tag{9}$$

and

$$\langle V \rangle_1 = \langle V \rangle_0 - c \tag{10}$$

Linear interaction energy models are based on the assumption that ΔF is proportional to $\langle V \rangle_1$:[2]

$$\Delta F = \alpha \langle V \rangle_1 \tag{11}$$

It is therefore of interest to compute, under the assumption of linear response, the proportionality coefficient α given by the ratio of ΔF to $\langle V \rangle_1$. From eqs 9 and 10

$$\frac{\Delta F}{\langle V \rangle_1} = \frac{\langle V \rangle_0 - c/2}{\langle V \rangle_0 - c} \tag{12}$$

The limiting values for this ratio are

$$\frac{\Delta F}{\langle V \rangle_1} = \begin{bmatrix} 1/2 \text{ if } |\langle V \rangle_0| \ll c \\ 1 \text{ if } |\langle V \rangle_0| \gg c \end{bmatrix} \tag{13}$$

Thus, under the assumption of linear response, the limiting values of the ratios between the free energy of adding solute-environment interactions and the corresponding average interaction energy are 1/2 and 1. The free energy change is half the interaction energy when the fluctuation of V, measured by $c = \langle (\delta V)^2 \rangle / kT$, is much larger than $\langle V \rangle_0$, the mean ligand-environment interaction energy calculated within the ensemble of conformations obtained in the absence of ligand-environment interactions. In the opposite limit at which the fluctuations of V are much smaller than $\langle V \rangle_0$, eq 12 gives $\Delta F / \langle V \rangle_1 = 1$, that is the free energy change is equal to the average solute-environment interaction energy.

In the following we apply linear response to the problem of binding free energy estimation and use eq 12 and physical arguments to derive values of the linear response coefficients. We will first review the derivation in explicit solvent and then examine the case in which the solvent is treated implicitly.

2.2. LIE Models in Explicit Solvent. *2.2.1. Hydration Free Energy − Explicit Solvent.* To successfully apply linear response ideas to the hydration free energy estimation, the process of hydration is described as occurring in stages. First the solute cavity is formed in the solvent, then solute−solvent van der Waals interactions are turned on, and finally solute−solvent electrostatic interactions are established. The process of cavity formation is dominated by excluded volume effects and solvent reorganization and does not conform well to the linear response formalism. When using a hard-sphere cavity interaction potential, the solute−solvent interaction energy is zero when the solute and the solvent do not overlap, and it is infinite when overlaps occur. In this limit the average solute−solvent interaction energy is identically zero because conformations in which solute cavity-solvent overlaps exist do not appear in the ensemble. Computational studies of cavity formation have generally been conducted using a continuous but sharp solute−solvent repulsive interaction potential.[23] In these cases, however, due to strong nonlinearities in the response of the solvent to the introduction of the solute cavity, the cavity hydration free energy is poorly correlated with the average repulsive cavity interaction potential. For cavity hydration free energy estimation a term proportional to the solute surface has been shown in some cases to be a reasonably good estimator for the free energy of cavity formation in water.[24,25] The proportionality coefficient γ between the cavity hydration free energy and the surface area can be interpreted as a surface tension coefficient. Indeed molecular simulations have obtained values of this proportionality coefficient similar to the value of the experimental air−water surface tension coefficient.[23,26]

Assuming linear response for the processes of introducing van der Waals and electrostatic solute−solvent interactions, the free energy change at each stage is proportional to the appropriate solute−solvent interaction energy (the solute−solvent van der Waals interaction energy and the solute−solvent electrostatic interaction energy, respectively) averaged in the state corresponding to the end of each stage (the uncharged solute and the fully interacting solute, respectively). The linear response coefficients are given by eq 12 and are, in general, different for each stage. Finally, making the approximation that all averages (denoted by $\langle \cdots \rangle_w$) are calculated when the solute fully interacts with the solvent, a LIE model for the hydration free energy is obtained[27]

$$\Delta F_h \simeq \alpha \langle V_{vdw} \rangle_w + \beta \langle V_{el} \rangle_w + \gamma \langle A \rangle_w \tag{14}$$

where α and β are the linear response proportionality coefficients for the van der Waals and electrostatic stages of the hydration process, V_{vdw} and V_{el} are the solute−solvent van der Waals and electrostatic interaction energies, respectively, γ is an empirical surface tension coefficient, and A is the solvent accessible surface area of the solute. As we show below, the form of the cavity hydration term can be justified in terms of linear response when the solvent is modeled implicitly.

Regression equations based on eq 14 have been parametrized against known experimental hydration free energies of small molecules.[27] However, linear response theory provides a way to estimate some of these parameters from first principles. It has been observed that the charging free energy of ionic and polar solutes in water is approximately proportional to the average solute−solvent electrostatic interaction energy with a linear response proportionality coefficient β of 1/2. According to eq 13 this occurs when the solvent reaction field in the absence of solute charges is much smaller than the fluctuations of the solvent reaction field. These conditions have been indeed verified by numerical studies, confirming that in general water behaves as a good linear dielectric medium.[6,28,29] These observations have constituted the basis for electrostatic linear response free energy models for solutions[2,6,7] as well as for the success of continuum dielectric models of water.[28,30−35]

It is well-known that the process of adding solute−water van der Waals interactions has different characteristics than the process of adding electrostatic interactions. It has been shown that the free energy change for adding attractive van der Waals interactions can be well approximated by the average solute−water van der Waals interaction energy.[23,36,37]

This implies that in this case $\Delta F/\langle V_{vdw}\rangle_1 \simeq 1$, which, under the assumption of linear response occurs (see eq 13) when the mean solute−solvent attractive van der Waals interaction energy in the absence of solute−solvent van der Waals interactions, is larger than the variance of the same quantity divided by kT. This is due to the fact that, although the solvent responds linearly to the solute perturbation, this response is smaller than the attractive van der Waals solvent field that exists at the solute location in the absence of solute−solvent van der Waals interactions. Contrary to dipolar fields in water that tend to cancel each other in the absence of a polarization source, van der Waals interactions are always additive. Based on this analysis we conclude that the α coefficient in the LIE regression equation should assume a value near 1. Explicit solvent simulations have generally confirmed this theoretical prediction.[23] Carlson and Jorgensen[27] have instead reported that a value of α significantly smaller than 1 is obtained by fitting the LIE equation to experimental hydration free energies of small molecules. We believe that the small value of α obtained by Carlson and Jorgensen is caused by compensation between the van der Waals and surface area fitting coefficients. The correct relative magnitude of these two coefficients is difficult to pinpoint because they correspond to highly correlated descriptors (the van der Waals solute−solvent interaction energy and the solute surface area). Indeed both the α and γ coefficients obtained by Carlson and Jorgensen are smaller than well established theoretical predictions.[24,36,38] We have reanalyzed the data from Tables 2 and 3 of ref 27. By setting $\alpha = 1$ and allowing for a nonzero intercept (to fit relative hydration free energies rather than the absolute ones) we achieved a nearly equivalent fit to the experimental hydration free energies. Specifically, using the Coulombic+van der Waals+solvent-accessible surface area model of Carlson and Jorgensen we reproduce the parameters reported previously,[27] $\alpha = 0.49$, $\beta = 0.42$, and $\gamma = 20$ cal/mol Å^2, with a cross-validated correlation of $R_{pred}^2 = 0.83$, whereas our model with α set to 1 gives $\beta = 0.49$ and $\gamma = 62$ cal/mol Å^2 with $R_{pred}^2 = 0.81$. The RMSD from the experimental free energies of hydration of the two models are also similar, 0.88 and 0.98 kcal/mol, respectively. Furthermore, the value of γ (62 cal/mol Å^2) we obtained from the analysis of the data of Carlson and Jorgensen is closer to the experimental value of the macroscopic vacuum-water surface tension (104 cal/mol Å^2)[39] and is similar to the value of the microscopic surface tension parameter obtained from explicit solvent estimates of the work of cavity formation in water (73 cal/mol Å^2).[23] These observations indicate that the simulation data obtained by Carlson and Jorgensen is consistent with the linear response behavior for these solutes.

In conclusion, this analysis shows that eq 14 should provide a good approximation of hydration free energies with the following choice of LIE coefficients: $\alpha \simeq 1$, $\beta \simeq 1/2$, and $\gamma \simeq 73$ cal/mol Å^2, the previously reported explicit solvent estimate.[23]

2.2.2. Binding Free Energy Estimation − Explicit Solvent. The binding free energy ΔF_b of a ligand to a receptor is taken as the difference of the work ΔF_c of creating the ligand in the receptor and the work ΔF_h of creating the ligand in solution. For the process of creating the ligand in the receptor an expression similar to eq 14 has been proposed[3,9] where the interaction energies include ligand−water interactions as well as ligand−receptor interactions, and A is taken as the solvent accessible surface area of the ligand in the receptor−ligand complex. By taking the difference between the LIE estimates of the insertion free energies in the receptor and in solution and assuming that the same values of the LIE coefficients α, β, and γ are appropriate for both, the following binding free energy LIE model is obtained

$$\Delta F_b = \Delta F_c - \Delta F_h \simeq \alpha(\langle V_{vdw}^c\rangle_c - \langle V_{vdw}^f\rangle_w) + \beta(\langle V_{el}^c\rangle_c - \langle V_{el}^f\rangle_w) + \gamma(\langle A\rangle_c - \langle A\rangle_w) \quad (15)$$

where V^c represents an interaction energy between the ligand and the environment (receptor and solvent), and V^f is the corresponding ligand−solvent interaction energy in the absence of the receptor (free ligand), A is the solvent accessible surface area of the ligand, $\langle\cdots\rangle_c$ represents an ensemble average with the ligand in the receptor pocket, and $\langle\cdots\rangle_w$ represents the corresponding average in solution.

Equation 15 and its variations are widely used in binding free energy prediction applications.[40−42] In these studies the LIE coefficients α, β, and γ are obtained by fitting eq 15 to known experimental binding free energies. In principle, linear response theory arguments could be applied, as for the case of hydration free energy estimation, to gain insights into the expected values of these coefficients. It is less clear, however, to what extent the receptor/solvent environment can be considered an ideal linear dielectric medium[10,43] and what value of the proportionality coefficient to use to estimate the electrostatic charging free energy from the ligand-environment electrostatic interaction energy. The LIE equation eq 15 implicitly assumes that the same electrostatic proportionality coefficient, β, applies to both the charging process in solution and in the protein receptor. However, contrary to the water environment, many proteins produce strongly anisotropic electrostatic fields. It is therefore reasonable to assume that a non-negligible electrostatic field exists at the binding site even in the absence of ligand charges. Under the assumption of linear response, eq 12 indicates that the presence of a residual average electrostatic potential at the ligand charge sites in the absence of ligand charges (that is $\langle V_{el}\rangle_0 \neq 0$) will cause the ratio $\Delta F/\langle V_{el}\rangle_1$ to deviate from the ideal solution value of 1/2. This analysis predicts that, although assuming $\beta = 1/2$ is a reasonable first guess, in general it would be advantageous to adopt a LIE regression equation in which the electrostatic estimator is split into a receptor environment component and a solution environment component each multiplied by an independent LIE coefficient.[44] Similarly, the LIE eq 15 implicitly assumes that the work of cavity formation in the protein receptor can be also estimated by the ligand surface area using a single surface tension coefficient applicable to both the water and receptor environments. However, due to the complex reorganization of the receptor binding pocket induced by ligand binding, the solute surface area is likely a poor descriptor for the free energy of ligand cavity formation in protein receptors. In this work we explore the alternative approach

of modeling the work of inserting the ligand in the receptor independently from the work of inserting the ligand in solution.

Most of the limitations of the LIE regression equation underlined here are mitigated in practice by cancellation of errors. The main goal of LIE studies is to obtain relative binding free energies within a family of related ligands, rather than absolute binding free energies. The overall deviation of estimated absolute binding free energies from the experimental affinities is accounted for by an adjustable intercept parameter that is often added to the LIE regression equation (see eq 15).[44] This intercept parameter obviously does not affect relative binding free energies estimated from the LIE equation. For example, the differences of cavity formation free energies between ligands of similar shape could be well represented by a surface area descriptor even though the individual absolute values are not. In this context the limits of the LIE regression equation are manifested in the limited range of applicability of a particular parametrization rather than in the accuracy of the LIE parametrization for a particular ligand set. A better understanding of the properties of the LIE equation could therefore lead to improved LIE ligand coverage and to rules that, based on properties of the ligand, associate a particular parametrization to a particular class of ligands.

2.3. The AGBNP Implicit Solvent Model. The Analytical Generalized Born plus Non-Polar (AGBNP) implicit solvent model[35] is based on an analytical pairwise descreening implementation of the Generalized Born model and a nonpolar hydration free energy model consisting of an estimator for the solute−solvent van der Waals dispersion energy and a surface area term corresponding to the work of cavity formation.

In the Generalized Born (GB) model[33] the electrostatic component of the hydration free energy is estimated as

$$G_{\mathrm{el}} \simeq G_{\mathrm{GB}} = -\frac{1}{2}\left(\frac{1}{\epsilon_{\mathrm{in}}} - \frac{1}{\epsilon_{\mathrm{w}}}\right)\sum_{ij}\frac{q_i q_j}{f_{ij}} \tag{16}$$

where ϵ_{in} is the dielectric constant of the interior of the solute, ϵ_{w} is the dielectric constant of the solvent (in this work $\epsilon_{\mathrm{in}} = 1$ and $\epsilon_{\mathrm{w}} = 80$), q_i and q_j are the charges of atom i and j, and

$$f_{ij} = \sqrt{r_{ij}^{\,2} + B_i B_j \exp(-r_{ij}^{\,2}/4B_iB_j)} \tag{17}$$

where r_{ij} is the distance between atoms i and j, and B_i and B_j are the Born radii of atoms i and j defined below. The summation in eq 16 runs for all atom pairs (i, j) including $i = j$. The diagonal $i = j$ terms can be separated from off-diagonal terms $i \neq j$ yielding the equivalent expression

$$G_{\mathrm{GB}} = -\frac{1}{2}\left(\frac{1}{\epsilon_{\mathrm{in}}} - \frac{1}{\epsilon_{\mathrm{w}}}\right)\left(\sum_i \frac{q_i^{\,2}}{B_i} + 2\sum_{i<j}\frac{q_i q_j}{f_{ij}}\right) \tag{18}$$

The first summation at the right-hand side of eq 18 is the sum of the GB self-energies of the atoms of the molecule, and the second term is the sum of the GB pair-energies. The self-energy of atom i corresponds to the solvation energy of the solute when only the charge of atom i is nonzero. It measures the energy of atom i in the reaction field due to the polarization of the solvent induced by the partial charge of atom i in the solute cavity. The self-energy is largest for the atoms that are most exposed to the solvent because they are capable of inducing stronger polarization fields. This effect is captured by the GB model in that atoms exposed to the solvent have smaller Born radii, whereas buried atoms tend to have larger Born radii. The pair-energy term corresponds to the dampening of electrostatic interactions in a high dielectric medium due to the screening of the solute charges. The GB equation (eq 18) can be shown to be an exact representation of the electrostatic charging free energy of the solute in a continuum dielectric in the two limiting cases of infinite atomic separation and complete atomic overlap.[45]

The Born radius of atom i is defined as the radius of the monatomic solute with partial charge q_i whose continuum dielectric hydration free energy is equal to the self-energy of atom i. In the Coulomb field approximation,[46,47] the Born radius is expressed as an integral centered on the position $\mathbf{r}_i$ of atom i

$$\frac{1}{B_i} = \frac{1}{R_i} - \frac{1}{4\pi}\int_{\Omega i} d^3\mathbf{r}\,\frac{1}{(\mathbf{r} - \mathbf{r}_i)^4} \tag{19}$$

where Ω_i is the bounded region corresponding to the solute volume excluding the atomic sphere corresponding to atom i, and R_i is the van der Waals radius of atom i. $1/R_i$ is the inverse Born radius of atom i in the absence of all the other solute atoms. The second term on the right-hand side of eq 19 takes into account the displacement of the solvent dielectric due to the other solute atoms. In pairwise solute descreening schemes this term is approximated by a pairwise sum[48,49] over the volumes of the neighboring atoms, which are traditionally empirically adjusted to account for atomic overlaps. The AGBNP model instead makes use of a parameter-free geometrical algorithm to calculate the volume scaling coefficients used in the pairwise descreening scheme. The same algorithm is also used to calculate atomic surface areas. This feature is particularly advantageous in ligand binding applications when parametrizations of volume scaling coefficients are not available for some chemical groups. It has been shown that AGBNP gives excellent agreement for the GB self-energies and surface areas in comparison to accurate, but much more expensive, numerical evaluations.[35]

The nonpolar model adopted in this work differs from most other implicit hydration free energy models in that the nonpolar component G_{np} of the hydration free energy is subdivided into cavity and solute−solvent van der Waals interaction terms

$$G_{\mathrm{np}} = G_{\mathrm{cav}} + G_{\mathrm{vdw}} \tag{20}$$

rather than estimated as a whole using a surface area model. The cavity component is described by a surface area model[23−26]

$$G_{\mathrm{cav}} = \sum_i \gamma_i A_i \tag{21}$$

where the summation runs over solute atoms, A_i is the van der Waals surface area of atom i, and γ_i is the surface tension parameter assigned to atom i.[35] The solute−solvent van der Waals free energy term is modeled by the expression

$$G_{vdw} = \sum_i \alpha_i \frac{a_i}{(B_i + R_w)^3} \quad (22)$$

where α_i is an adjustable dimensionless atomic parameter on the order of 1,[35] B_i is the Born radius of atom i, $R_w = 1.4$ Å is the radius of a water molecule, and

$$a_i = -\frac{16}{3}\pi\rho_w\epsilon_{iw}\sigma_{iw}^6 \quad (23)$$

where $\rho_w = 0.033428$ Å^{-3} is the number density of water at standard conditions, and σ_{iw} and ϵ_{iw} are the OPLS force field[50] Lennard-Jones interaction parameters for the interaction of solute atom i with the oxygen atom of the TIP4P water model.[51] Equation 22 is derived by integrating the attractive component of the ligand−water Lennard-Jones potential over the solvent volume assuming homogeneous solvent density.[35]

2.4. LIE Models with Implicit Solvation. *2.4.1. Hydration Free Energy − Implicit Solvent.* If the solvent is described explicitly, the energetic descriptors in eq 14 are evaluated simply by averaging the sum of pair interaction energies between ligand atoms and solvent atoms. It is of interest to derive the expressions for the corresponding estimators when the solvent is treated implicitly. In this case the expression for the canonical configurational partition function of the system includes explicitly only solute degrees of freedom, r^N, and the solute−solvent and solvent−solvent potential energies are replaced by the solvent potential of mean force $W(r^N)$[32,52]

$$Z(\lambda) = \int dr^N \exp[-u(r^N)/kT] \exp[-\lambda W(r^N)/kT] \quad (24)$$

where $u(r^N)$ is the intramolecular potential energy of the solute. The original system is replaced by an equivalent reduced system characterized by the effective potential energy function $U_{eff} = u(r^N) + W(r^N)$. In this section we derive expressions for the LIE estimators in implicit solvent that, based on linear response theory, best correspond to their explicit solvent counterparts. The main difference between the explicit and implicit solvent representations is that the latter lacks fluctuations due to solvent motion. We will show that the response of the implicit solvent environment is significantly different from the response of the explicit solvent environment.

In this study we employ a solvent potential of mean force for water of the form[35]

$$W(r^N) = G_{el} + G_{vdw} + G_{cav} \quad (25)$$

where, as described in the previous section, G_{el} is the electrostatic component (modeled in this work using the Generalized Born model), G_{vdw} corresponds to solute−solvent attractive van der Waals interactions, and G_{cav} is the work for creating the solute cavity in the solvent estimated using a model based on the solute surface area. The process of inserting the solute into the solution can be thought of, starting with $W(r^N) = 0$, as first turning on the cavity component G_{cav}, then the van der Waals component G_{vdw}, and finally the electrostatic component G_{el}. This is analogous to the process of turning on explicit solute−solvent interactions described in the previous section. For example the process of adding the electrostatic component G_{el} after having already added the van der Waals and cavity components (G_{vdw} and G_{cav}) is described by the λ-dependent effective potential energy of eq 2 where

$$U_0 = u + G_{vdw} + G_{cav} \quad (26)$$

is the reference potential, and

$$V = G_{el} \quad (27)$$

is the perturbation. It follows that the linear response estimator for the process of adding solute−implicit solvent interactions is $\langle G_{el}\rangle_w$, the average of the electrostatic implicit solvent term in the ensemble of solute conformations generated when G_{el} is turned on.

Assuming linear response, the ratio between the free energy change for this process and the estimator $\langle G_{el}\rangle_w$, which determines the expected ideal value for the corresponding LIE coefficient, is given by eq 12. In this case, unlike the corresponding electrostatic explicit solvent Coulombic term, the average of G_{el}, $\langle G_{el}\rangle_0$, over the ensemble of conformations obtained in absence of G_{el} in general is not small. Moreover, due to the lack of contributions from solvent motion, the variance, $\langle(\delta G_{el})^2\rangle_0$, of G_{el} is expected to be smaller in relation to the variance of the solute−solvent electrostatic interaction energy in explicit solvent. Equation 12 indicates that, due to both of these effects, the expected value of the electrostatic LIE regression coefficient β when the solvent is treated implicitly should be closer to 1 rather than 1/2. This conclusion applies to the cavity and van der Waals implicit solvent estimators as well.

This is best appreciated in the limiting case when the solute is treated rigidly. In this case, because there are no variable degrees of freedom, $\langle(\delta G_{el})^2\rangle_0 = 0$, and the ensemble average of G_{el} reduces to the value of G_{el} for the given solute conformation. This is true for any of the solvation energy components (cavity, van der Waals, and electrostatic). Therefore, insofar as the solute can be treated as a rigid molecule and the implicit solvent model gives an accurate estimate of the electrostatic hydration free energy of the solute, we have

$$\Delta F_h \simeq G_{el} + G_{vdw} + G_{cav} \quad (28)$$

which is the LIE regression equation for the hydration free energy with all of the adjustable coefficients set to 1. In practice due to solute flexibility and limitations of the implicit solvent model, a LIE regression equation which includes adjustable LIE coefficients fit to experimental hydration free energies

$$\Delta F_h \simeq \alpha\langle G_{vdw}\rangle_w + \beta\langle G_{el}\rangle_w + \gamma\langle G_{cav}\rangle_w \quad (29)$$

which is equivalent to eq 14 when the free energy of cavity formation is modeled using the solute surface area, is

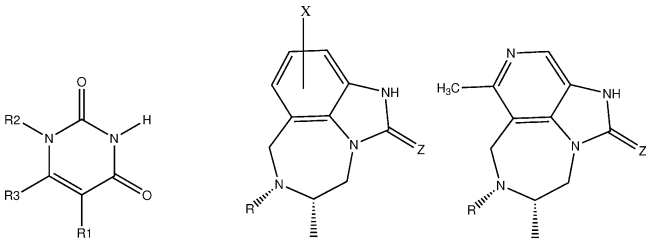

Figure 1. The core structures of the HEPT and TIBO analogues. The TIBO core to the right corresponds to the TIBO pyridyl analogues 27a, 27b, 27c, and 27d.

expected to yield more accurate predictions. Nevertheless the values of the adjustable coefficients α, β, and γ in eq 29 are expected to assume values near 1.

2.4.2. Ligand Binding Free Energy − Implicit Solvent. As pointed out in the previous section, the binding free energy of a ligand to a receptor is calculated as the difference between the work of inserting the ligand in the receptor site solvated by water and the work of inserting the ligand molecule in solution in the absence of the receptor. The LIE equation for the work of creating the ligand in the receptor site takes the same form as eq 29, where now the expressions for the estimators should take into account the fact that the ligand interacts implicitly with the solvent, as modeled by the solvent potential of mean force, as well as with the receptor atoms which are treated explicitly.[10,12] Here we determine, based on linear response theory, the relationship between the free energy changes and the corresponding average effective potential energy changes when part of the system is treated explicitly and part is treated implicitly using the AGBNP implicit solvent model. This is complicated by the fact that implicit solvent models are in general nonpair decomposable. That is it is not possible to define a solute−implicit solvent interaction energy based on pairwise sums between explicit atoms. In the Appendix we derive the following expression for the LIE regression equation for the free energy for creating the ligand in the receptor site

$$\Delta F_c \simeq \alpha \langle V_{LJ} + G^c_{vdw} - G^p_{vdw}\rangle_c + \beta \langle V_{el} + 2(G^c_{el} - G^p_{el})\rangle_c + \gamma \langle G^c_{cav} - G^p_{cav}\rangle_c \quad (30)$$

where all of the averages $\langle\cdots\rangle_c$ are taken for the complex with the ligand fully interacting with the receptor atoms and the solvent continuum. V_{LJ} is the ligand−receptor Lennard-Jones interaction energy, V_{el} is the ligand−receptor electrostatic interaction energy, and G^c and G^p represent implicit solvent free energy terms of the complex and receptor, respectively. Quantities such as $G^c_{el} - G^p_{el}$ are the difference between each implicit solvent energy term evaluated for the complex conformation and the same conformation without the ligand.

The form of the electrostatic LIE estimator for the insertion of the ligand in the complex, $\langle V_{el} + 2(G^c_{el} - G^p_{el})\rangle$, that emerges from our derivation is similar to that proposed by Carlsson et al.[12] Ours includes ligand−receptor interactions (ligand−receptor Coulomb interactions and the pair GB interaction energy between ligand atoms and receptor atoms) and ligand properties (the GB self-energy of the ligand and the GB pair energy between ligand atoms) as well as receptor properties (the change of the GB self-energy of receptor atoms and GB pair energy between receptor atoms due to the displacement of the solvent dielectric by the ligand), which are not included in the electrostatic estimator of Carlsson et al.[12] The implicit solvent LIE van der Waals estimator $\langle V_{LJ} + (G^c_{vdw} - G^p_{vdw})\rangle$ includes ligand−receptor Lennard-Jones interactions as well as two new terms: the ligand−implicit solvent van der Waals interaction energy and the change in the receptor−implicit solvent van der Waals interaction energy upon ligand complexation. This is in contrast to the corresponding estimator in explicit solvent which includes the interactions of the ligand with the receptor and the solvent. This difference can be understood in terms of the additional coupling between ligand atoms and receptor atoms introduced by the partial averaging inherent in the definition of the solvent potential of mean force. In the explicit solvent case the LIE estimator includes explicitly ligand−solvent interactions. In the implicit solvent case the mean effect of the solvent is replaced by effective ligand−receptor as well as ligand−ligand and receptor−receptor interactions.

As previously noted,[12] the implicit LIE solvent electrostatic estimator used earlier by Zhou et al.[4] lacks receptor desolvation contributions. They proposed an electrostatic implicit solvent LIE estimator composed of the GB self-energies of the ligand atoms, the GB pair energies between ligand atoms, and half the GB pair energies between ligand atoms and receptor atoms. The electrostatic estimator derived here based on linear response analysis includes the full amount of the GB pair energies between ligand atoms and receptor atoms and, in addition, the change of self-energies and GB pair energies of the receptor atoms due to the introduction of the ligand. The major difference between the two implementations is the absence of receptor desolvation contributions in the model of Zhou et al. Furthermore, our van der Waals estimator, $\langle V_{LJ} + (G^c_{vdw} - G^p_{vdw})\rangle$, includes estimates of the loss of ligand−solvent and receptor−solvent van der Waals interactions that are not explicitly included in the implicit solvent model adopted by Zhou et al.[4] and Carlsson et al.[12]

To obtain a LIE regression equation for the binding free energy ΔF_b, eqs 30 and 29 can be combined in at least two ways. The first follows previously proposed LIE models in both explicit[3] and implicit[4] solvents. In these models a LIE regression equation is obtained by subtracting eq 29 from eq 30, assuming that the values of LIE coefficients α, β, and γ are the same for the processes of inserting the ligand in solution and in the receptor environment

$$\Delta F_b \simeq \alpha[\langle V_{LJ} + (G^c_{vdw} - G^p_{vdw})\rangle_c - \langle G^f_{vdw}\rangle_w] + \beta[\langle V_{el} + 2(G^c_{el} - G^p_{el})\rangle_c - \langle G^f_{el}\rangle_w] + \gamma[\langle G^c_{cav} - G^p_{cav}\rangle_c - \langle G^f_{cav}\rangle_w] \quad (31)$$

where G^f denotes implicit solvent terms for the ligand free in solution. Alternatively the processes of ligand formation in the receptor and in solution can be decoupled by considering the LIE coefficients for the process of insertion of the ligand in the receptor (eq 30) as indepndent adjustable parameters, whereas the hydration free energy is assumed proportional to the implicit solvent estimate (eq 28) with a

Table 1: Molecular Structures and Experimental IC_{50} and Resulting Binding Free Energies, ΔG_b of the HEPT Analogs

compd	IC_{50}[a]	ΔG_b[b]	R_1	R_2	R_3
H06	0.0027	−12.16	i-Pr	CH_2OCH_2Ph	SPh
H17	0.0027	−12.16	i-Pr	$CH_2OCH_2CH_2OH$	SPh-3,5-di-Me
H11	0.004	−11.89	i-Pr	$CH_2OCH_2CH_3$	CH_2Ph
H18	0.0059	−11.68	Et	CH_2OCH_2Ph	SPh
H10	0.012	−11.24	i-Pr	$CH_2OCH_2CH_3$	SPh
H16	0.013	−11.19	Et	$CH_2OCH_2CH_2OH$	SPh-3,5-di-Me
H09	0.019	−10.96	Et	$CH_2OCH_2CH_3$	SPh
H05	0.088	−10.01	Me	CH_2OCH_2Ph	SPh
H12	0.1	−9.93	c-Pr	$CH_2OCH_2CH_3$	SPh
H15	0.26	−9.35	Me	$CH_2OCH_2CH_2OH$	SPh-3,5-di-Me
H03	0.33	−9.20	Me	$CH_2OCH_2CH_3$	SPh
H20	1.2	−8.40	Me	Bu	SPh
H04	2.1	−8.06	Me	CH_2OCH_3	SPh
H07	2.2	−8.03	Me	Et	SPh
H02	3.6	−7.73	Me	$CH_2OCH_2CH_2CH_3$	SPh
H01	7.0	−7.32	Me	$CH_2OCH_2CH_2OH$	SPh
H13	23.0	−6.52	Me	$CH_2OCH_2CH_2OH$	CH_2Ph
H14	85.6	−5.78	Me	$CH_2OCH_2CH_2OH$	OPh
H08	150[c]	−5.43	Me	Me	SPh
H19	250[c]	−5.11	Me	H	SPh

[a] From ref 53, in μM units. [b] $kT \ln IC_{50}$ at $T = 310$ K, in kcal/mol. [c] Largest concentration tested.

proportionality coefficient ω:

$$\Delta F_b \simeq \alpha\langle V_{LJ} + (G^c_{vdw} - G^p_{vdw})\rangle_c + \beta\langle V_{el} + 2(G^c_{el} - G^p_{el})\rangle_c + \gamma\langle G^c_{cav} - G^p_{cav}\rangle_c - \omega(\langle G^f_{vdw}\rangle_w + \langle G_{el}{}^f\rangle_w + \langle G^f_{cav}\rangle_w) \quad (32)$$

In this work both of these approaches are considered.

3. Ligand Sets and Binding Free Energies

Figure 1 and Tables 1 and 2 show the molecular structures of the HEPT and TIBO analogues. The HEPT NNRTI ligand set used here is the same as the one in the LIE studies by Rizzo et al.[53] in explicit solvent and Zhou et al.[4] in implicit solvent. It is included here for comparison with previous studies. The TIBO NNRTI ligand set has been compiled from the work of Ho et al.[54] who reported activities (IC_{50}'s) of 40 TIBO derivatives and pyridyl analogues. We have included in our study all of the compounds reported by Ho et al. (with the exception of those few for which only an activity range was reported rather than a single activity value) plus additional compounds which were part of a TIBO subset studied previously by LIE modeling.[55] The resulting 40 TIBO ligands which were simulated in our study are listed in Table 2. (Based on the values of the energetic descriptors obtained from the simulation, 3 of these 40 ligands were not included in the LIE training set, see below.) HEPT compounds are labeled using the naming scheme in ref 53; for the TIBO compounds we adopted the names in refs 54 and 55. In Table 1 H11 is the reference HEPT compound MKC-442, and in Table 2 the reference compound is 1a (also known as TIBO-9Cl).

4. Simulation Procedure

A model of the NNRTI binding site is constructed as previously described[53] by including in the simulation only the region of HIV-RT closest to the ligand. The same set of 124 residues was included in the model for both the TIBO and HEPT compounds. These 124 residues are organized in 7 protein segments: residues 91−110, 161−205, 222−242, 316−321, 343−349, and 381−383 in the p66 subunit and 134−140 in the p51 subunit. The end of each fragment is capped with acetyl (ACE) or *N*-methylamide (NMA) groups. Residues 95−108, 179−183, 186−191, 198, 225−229, and 318−319 of the p66 subunit and residues 136 and 138 in the p51 subunit are free to move in the MD simulation. The positions of the atoms of residues 94, 109, 178, 184, 197, 199, 224, 230, 231, and 240 in the p66 subunit and 135, 137, and 139 in the p51 subunit are restrained using a harmonic restraining potential with a force constant of 25 kcal/mol Å^2. All other residues are held fixed. Residues LYS101, LYS102, LYS103, LYS104, and LYS238 are protonated, whereas ASP186, ASP192, ASP237, GLU138, and GLU233 are deprotonated. All other ionizable residues are set in their neutral charge state.

We employ the structure of the complex of HIV-1 RT with MKC-442 (PDB id 1rt1)[56] as the template structure from which the initial structures of all the HEPT complexes are constructed. For the TIBO compounds the corresponding template structure is from the complex with 9-Cl TIBO[57] (PDB 1tvr). The template structures were first energy minimized, and then each complex was constructed by modifying the ligand starting from the corresponding reference ligands (H11 and 1a in Tables 1 and 2) by manual editing of the ligand structure using Maestro.[58] Each complex as modified is then energy minimized to relieve steric clashes. One conformation of each ligand was constructed. The ligand conformations thus obtained were also used as starting conformations for the molecular dynamics simulations in the solution phase using the same protonation state and force field parameters as in the receptor-bound simulations.

All molecular dynamics calculations are performed using the program IMPACT[59] with the 2001 parametrization of the OPLS-AA/AGBNP force field.[35,50] The RESPA multiple

Table 2: Molecular Structures and Experimental IC_{50} and Resulting Binding Free Energies, ΔG_b of the TIBO Analogs

compd	IC_{50}[a]	ΔG_b[b]	X	R	Z
10f	0.0030	−12.09	8-Br	DMA	S
6a	0.0043	−11.87	8-Cl	DMA	S
6c	0.0050	−11.78	8-SCH_3	DMA	S
6b	0.0058	−11.69	8-F	DMA	S
10m	0.0136	−11.16	8-CH_3	DMA	S
17	0.0250	−10.79	9-F	DMA	S
20	0.0255	−10.76	9,10-di-Cl	DMA	S
1d	0.0295	−10.69	8-CCH	DMA	S
10j	0.0296	−10.68	8-CH_2CH_3	DMA	S
1a	0.034	−10.6	9-Cl	DMA	S
6d	0.034	−10.6	8-OCH_3	DMA	O
1	0.044	−10.44	H	DMA	S
10h	0.0473	−10.39	8-I	DMA	S
10e	0.0474	−10.39	8-Br	DMA	O
10b	0.0563	−10.29	8-CN	DMA	S
10g	0.088	−10.01	8-I	DMA	O
6e	0.0959	−9.96	8-OCH_2CH_3	DMA	S
10c	0.188	−9.54	8-COH	DMA	S
27c	0.243	−9.39		DEA	S
18c	0.3142	−9.23	8-CH_3	DEA	O
1i	0.4371	−9.02	8-CCH	DMA	O
10i	0.4376	−9.02	8-CH_2CH_3	DMA	O
18b	0.485	−8.96	9-CF_3	DMA	S
10l	0.989	−8.52	8-CH_3	DMA	O
21	1.075	−8.47	10-Br	DMA	S
10a	1.1396	−8.43	8-CN	DMA	O
27b	2.0	−8.09		DEA	O
16	2.45	−7.96	9-NO_2	CPM	S
1l	3.155	−7.81	H	DMA	O
19b	4.725	−7.56	10-OCH_3	DMA	S
18a	5.919	−7.42	9-CF_3	DMA	O
19a	6.63	−7.35	10-OCH_3	DMA	O
14b	6.65	−7.35	8-$N(CH_3)_2$	CPM	O
15b	6.65	−7.35	9-$N(CH_3)_2$	CPM	O
13	33.43	−6.35	9-NO_2	CPM	O
15a	60.55	−5.98	9-NH_2	CPM	O
15c	159	−5.39	9-$NHCOCH_3$	CPM	O
27d	596	−4.58		CPM	O
14a	849	−4.36	8-NH_2	CPM	O
27a	872	−4.34		DMA	O

[a] From ref 54, in μM units. [b] $kT \ln IC_{50}$ at $T = 310$ K, in kcal/mol.

time step MD integrator with an inner time step of 0.25 fs for covalent interactions, and a time step of 1 fs for nonbonded interactions is employed. Temperature is controlled by velocity rescaling. The simulations for the unbound ligand consist of 10 ps of heating from 10 to 310 K and 25 ps of equilibration at 310 K followed by 100 ps of data collection. Simulations of the inhibitor−protein complex consist of 50 ps of heating from 10 to 310 K and 50 ps of equilibration at 310 K and 100 ps of data collection. A residue-based nonbonded neighbor list with a 15 Å distance cutoff was used in the inhibitor−protein complex simulations.

Energetic analysis is conducted on trajectory files collected during the data collection phase of each simulation. Each energetic quantity is averaged to obtain values of the LIE descriptors for each ligand according to eqs 33−40. Convergence was monitored by plotting the running average of each property.

5. Results

5.1. LIE Descriptors. The values of the calculated energetic descriptors for the two ligand sets are listed in Tables3 and 4. In these tables and in the rest of the paper one- and two-letters mnemonics are used to identify the calculated descriptors as follows

$$\mathrm{EC} = \langle V_{\mathrm{el}} \rangle_{\mathrm{c}} \tag{33}$$

$$\mathrm{ES} = \langle G^{\mathrm{c}}_{\mathrm{el}} - G^{\mathrm{p}}_{\mathrm{el}} \rangle_{\mathrm{c}} \tag{34}$$

$$\mathrm{EL} = \langle G^{\mathrm{f}}_{\mathrm{el}} \rangle_{\mathrm{w}} \tag{35}$$

$$\mathrm{LJ} = \langle V_{\mathrm{LJ}} \rangle_{\mathrm{c}} \tag{36}$$

$$\mathrm{V} = \langle G^{\mathrm{c}}_{\mathrm{vdw}} - G^{\mathrm{p}}_{\mathrm{vdw}} \rangle_{\mathrm{c}} \tag{37}$$

$$\mathrm{VL} = \langle G^{\mathrm{f}}_{\mathrm{vdw}} \rangle_{\mathrm{w}} \tag{38}$$

$$\mathrm{C} = \langle G^{\mathrm{c}}_{\mathrm{cav}} - G^{\mathrm{p}}_{\mathrm{cav}} \rangle_{\mathrm{c}} \tag{39}$$

$$\mathrm{CL} = \langle G^{\mathrm{f}}_{\mathrm{cav}} \rangle_{\mathrm{w}} \tag{40}$$

where V_{el} and V_{LJ} are, respectively, the Coulomb and Lennard-Jones interaction energies between the ligand and receptor atoms, G_{el}, G_{vdw}, and G_{cav} refer to, respectively, the electrostatic, van der Waals, and cavity components of the AGBNP implicit solvent model of either the receptor−ligand complex (denoted by "c"), the receptor (denoted by "p"), or the ligand (denoted by "f "). Averages are taken for either the receptor−ligand complex (denoted by $\langle \cdots \rangle_{\mathrm{c}}$) or the ligand free in solution (denoted by $\langle \cdots \rangle_{\mathrm{w}}$). The difference between complex and receptor energies are calculated for each conformation of the complex by calculating the energy of the complex and that of the receptor obtained by removing the ligand without changing the conformation of the receptor. For instance $G^{\mathrm{c}}_{\mathrm{el}} - G^{\mathrm{p}}_{\mathrm{el}}$ is the difference between the GB electrostatic energy of the given complex conformation and the corresponding quantity for the same conformation after removal of the ligand. See the Appendix for additional discussion of this point.

5.2. HEPT and TIBO Binding Modes. As was noted earlier, the HEPT inhibitors adopt a butterfly conformation as shown in Figure 3. The two wings are formed by the R_3 side chain (see Figure 1) and the ligand core. To accommodate the ligand the protein forms a complementary pocket which has been described as acting as a "shrink wrap".[60] Compared with the crystal structure of the complex with MKC-442, the molecular dynamics trajectories of the HEPT complexes show relatively little conformational variation. The aromatic side chain (R_3) interacts favorably with TYR181, TYR188, and TRP229, and the thiothymine ring interacts with LYS101 and LYS103. The mostly hydrophobic R_2 side chain contacts with LEU234, PHE227, and VAL106. The position of the R_3 side chain is well conserved for all ligands, indicating that π stacking between R_3 and TYR181 is a prerequisite for binding. The role of the aliphatic R_1

Table 3: Computed Values of the Energetic Descriptors for the HEPT Compounds[a]

compd	EC	ES	EL	LJ	V	VL	C	CL
H06	−12.59	−0.07	−20.35	−62.77	−2.97	−42.23	−17.93	37.37
H17	−14.05	3.72	−18.51	−58.92	−3.61	−39.82	−13.25	35.04
H11	−11.08	6.27	−10.63	−52.91	−2.95	−35.33	−11.04	31.32
H18	−16.85	6.03	−20.11	−62.79	−3.38	−40.87	−17.31	36.78
H10	−11.54	1.29	−16.55	−53.97	−2.54	−35.50	−11.89	31.51
H16	−12.03	3.02	−18.70	−59.93	−3.75	−38.59	−10.31	34.32
H09	−9.12	0.43	−16.36	−53.55	−3.58	−33.80	−11.20	30.34
H05	−11.94	1.47	−20.76	−59.83	−3.84	−39.59	−15.03	35.12
H12	−10.13	0.49	−15.67	−52.82	−2.63	−34.86	−11.58	31.39
H15	−11.93	3.23	−19.39	−53.99	−4.25	−37.59	−10.40	33.11
H03	−1.17	−6.10	−16.66	−49.15	−3.40	−33.15	−9.69	29.35
H20	−7.04	5.80	−8.38	−49.22	−2.72	−32.32	−8.90	28.19
H04	−13.67	5.39	−15.19	−50.67	−3.82	−33.53	−9.51	29.94
H07	−12.78	6.32	−9.46	−43.68	−2.50	−29.71	−9.00	26.16
H02	−11.21	3.00	−16.81	−50.49	−2.41	−35.08	−10.51	31.00
H01	−19.31	10.15	−18.38	−53.59	−2.25	−33.74	−12.09	29.88
H13	−16.09	5.17	−15.32	−42.51	−3.63	−31.09	−5.79	27.61
H14	−9.36	3.42	−19.14	−49.16	−3.02	−32.68	−11.26	29.04
H08	−14.38	8.75	−10.28	−39.42	−3.24	−28.18	−4.78	25.22
H19	−6.01	4.33	−10.18	−39.57	−2.16	−26.58	−7.58	24.42

[a] Values are in kcal/mol. The list is ordered from strongest to weakest binders.

side chain is to modulate the strength of the binding. Binding is favored by small branched groups in this position (see Table 1). As in previous simulations,[19] for most HEPT complexes we observe a stable hydrogen bond between the NH group of the thymine ring with the CO group of the LYS101 backbone. An additional hydrogen bond between the CO of the thymine ring and the NH group of LYS101 is also observed in some conformations collected from the simulations. Previous studies of HEPT in explicit solvent[53] have shown that the R_2 side chain is partly solvent exposed and forms hydrogen bonds with water molecules. Solvent exposure of this ligand side chain, which helps lower the desolvation penalty of these ligands, is also observed in our simulations in implicit solvent. For reasons that are not immediately apparent, the observed binding mode of the H03 HEPT ligand in the simulation differs from all the other HEPT ligands. Relative to the consensus binding mode the core of H03 is twisted, the hydrogen bond with the lysine residues is absent, and the R_2 side chain is less solvent exposed. This causes the unusually small values of the EC and ES descriptors of H03 (Table 3). Nevertheless most of the LIE models we developed (see below) predict binding free energies for H03 in agreement with the experimental binding free energy.

The binding mode of TIBO compounds (see Figure 3) is similar to that of the HEPT compounds except that more variability is observed. The dimthylallyl (DMA) or cyclopropylmethyl (CPM) substitutions in the R position of the TIBO cores (see Figure 1) replaces the aromatic side chain of the HEPT compounds in the R_3 position. As for the HEPT compounds, this group interacts with TYR181, TYR188, and TRP229 but without the extensive π stacking interactions characteristic of HEPT complexes. The DMA (or CPM) ligand side chain interacts with the protein in a variety of orientations and is not as constrained as the R_3 aromatic side chain of the HEPT compounds. In general we find that TIBO compounds form more hydrogen bond interactions with the receptor than the HEPT compounds. Both the NH group and the oxygen (or sulfur) atoms in the Z position of the TIBO cores form hydrogen bonds with the backbone of LYS101. For some ligands (6d, 18c, 1i, and 27b) we found an additional hydrogen bond between the carbonyl of the TIBO cores and the side chain of LYS103. These additional ligand−receptor hydrogen bonds are consistent with the larger ligand−receptor Coulomb interactions calculated for the TIBO compounds (compare the EC descriptor in Tables 3 and 4). Nonspecific electrostatic interactions, however, seem to play a role as well; although ligands 14a and 14b are the ones with the strongest ligand−receptor interaction energies, these ligands do not present on average more hydrogen bonds than the other TIBO ligands. The unusually small ligand−receptor electrostatic energy of ligand 6c is consistent with the absence for this ligand of one of the hydrogen bonds with LYS101. Ligand 10e also presents a weak electrostatic interaction with the protein. This ligand differs from all the other ligands in that it forms a single hydrogen bond with LYS103 rather than LYS101. The weak electrostatic interaction energies of these two ligands are counterbalanced by the correspondingly small receptor desolvation penalties (the ES descriptor in Table 4). In these complexes the LYS101 backbone remains partly solvent exposed suggesting the possibility of water-mediated interactions with the ligand. Both ligands 6c and 10e are characterized by stronger than average van der Waals interactions with the protein due to the almost parallel contact between the dimethylallyl group and the TYR181. Our simulations as well as similar observations by others[20] indicate that both for the HEPT and TIBO complexes a competition exists between hydrophobic and electrostatic interactions. The stronger the electrostatics interactions, the less the ligand is able to make good contacts with TYR181 and the other hydrophobic residues lining the binding pocket.

Table 4: Computed Values of the Energetic Descriptors for the TIBO Compounds[a]

compd	EC	ES	EL	LJ	V	VL	C	CL
10f	−24.44	13.60	−12.14	−51.75	−3.04	−35.46	−12.28	29.14
6a	−23.86	15.78	−9.10	−50.90	−1.85	−33.84	−14.33	28.36
6c	−12.44	9.32	−13.30	−56.02	−3.16	−36.28	−10.75	31.06
6b	−24.65	18.06	−9.68	−48.73	−1.34	−32.52	−14.43	28.40
10m	−24.69	18.06	−11.55	−51.28	−2.91	−34.00	−12.04	29.31
17	−22.29	13.80	−9.82	−47.91	−2.20	−32.22	−10.62	28.79
20	−29.38	19.59	−9.28	−51.51	−2.54	−36.83	−10.80	31.48
1d	−37.98	25.85	−10.63	−50.71	−2.75	−35.36	−11.94	30.16
10j	−24.22	16.04	−9.92	−49.67	−1.01	−33.53	−14.37	29.96
1a	−25.63	15.85	−9.64	−47.72	−2.73	−34.45	−8.65	29.67
6d	−47.70	28.26	−11.02	−48.21	−2.80	−34.58	−11.40	29.33
1	−23.83	15.04	−10.42	−45.50	−2.57	−31.81	−10.55	27.55
10h	−14.22	−3.20	−37.95	−50.13	2.04	−31.42	−11.13	31.10
10e	−10.86	4.07	−14.88	−48.42	−4.07	−34.78	−10.28	28.37
10b	−28.50	18.55	−11.25	−49.18	−1.79	−33.31	−10.77	29.71
10g	−23.80	3.27	−35.73	−48.53	2.68	−31.82	−13.77	31.71
6e	−35.59	24.21	−11.83	−52.48	−2.85	−36.35	−11.34	30.24
10c	−37.51	23.43	−14.57	−50.61	−2.23	−34.69	−11.10	30.84
27c	−21.95	14.29	−11.82	−54.00	−3.12	−36.90	−11.56	29.70
18c	−41.27	26.35	−12.77	−50.85	−2.96	−36.17	−14.27	30.02
1i	−49.47	29.96	−13.72	−49.11	−2.94	−34.44	−11.95	28.90
10i	−31.26	19.54	−12.87	−47.47	−1.95	−32.62	−11.35	28.84
18b	−29.20	20.78	−9.92	−48.65	−0.98	−34.13	−12.44	30.75
10l	−28.54	13.61	−15.33	−45.67	−2.78	−33.21	−10.20	28.31
21	−27.80	17.16	−13.30	−47.61	−3.81	−35.36	−8.19	29.06
10a	−41.51	24.46	−14.30	−47.44	−1.79	−32.75	−11.57	29.34
27b	−40.90	26.85	−12.48	−51.78	−3.20	−35.97	−11.29	29.60
16	−23.77	19.11	−10.07	−47.88	−1.91	−32.76	−9.20	29.32
1l	−25.13	13.89	−13.75	−45.98	−2.98	−31.04	−10.51	26.77
19b	−27.00	18.11	−11.24	−48.57	−3.50	−35.38	−9.757	30.82
18a	−37.30	24.09	−13.15	−45.89	−0.94	−33.39	−11.95	29.88
19a	−23.90	15.67	−14.68	−47.57	−3.11	−34.01	−12.42	29.13
14b	−46.70	25.80	−13.76	−45.08	−1.82	−33.33	−11.45	28.96
15b	−43.93	26.24	−13.72	−46.18	−2.22	−32.97	−11.03	27.95
13	−35.88	21.78	−14.49	−47.26	−1.46	−31.95	−11.56	28.35
15a	−23.75	16.16	−14.95	−45.27	−3.27	−29.81	−7.03	26.35
15c	−35.30	13.81	−30.85	−50.80	−3.66	−35.31	−9.00	30.61
27d	−38.32	22.49	−12.84	−43.96	−2.70	−30.56	−9.23	25.70
14a	−49.35	29.39	−14.70	−41.03	−1.34	−29.62	−12.22	25.74
27a	−41.09	26.34	−12.69	−47.97	−2.18	−32.58	−14.31	27.31

[a] Values are in kcal/mol. The list is ordered from strongest to weakest binders.

5.3. Comparison of HEPT and TIBO LIE Descriptors. The values of the calculated descriptors in Tables 3 and 4 show clear differences between the HEPT and TIBO data sets. The HEPT estimators based on van der Waals interactions (LJ, LJ + V and LJ + V − VL) exhibit stronger correlation with the experimental affinities than the corresponding TIBO estimators. A similar trend is observed for cavity estimators (C and C − CL). The van der Waals and cavity estimators both generally reflect the amount of surface area of the ligands. Indeed we find a strong linear correlation between the LJ and CL descriptors for the HEPT compounds ($R^2 = 0.94$). This correlation is significantly weaker for the TIBO compounds ($R^2 = 0.52$).

Due to the presence of additional ligand−receptor hydrogen bonds discussed above, the ligand−receptor electrostatic interaction energies (the EC descriptor) of the TIBO compounds are significantly more favorable than those of the

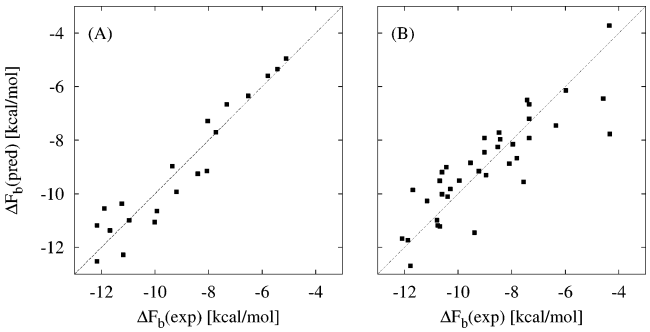

Figure 2. Predicted versus experimental binding free energies of the (A) HEPT and (B) TIBO analogues to HIV-RT for the MDL3 LIE model. The correlation coefficients and RMSDs for this model are 0.90 and 0.71 kcal/mol and 0.73 and 1.08 kcal/mol for the HEPT and TIBO compounds, respectively (see Table 5).

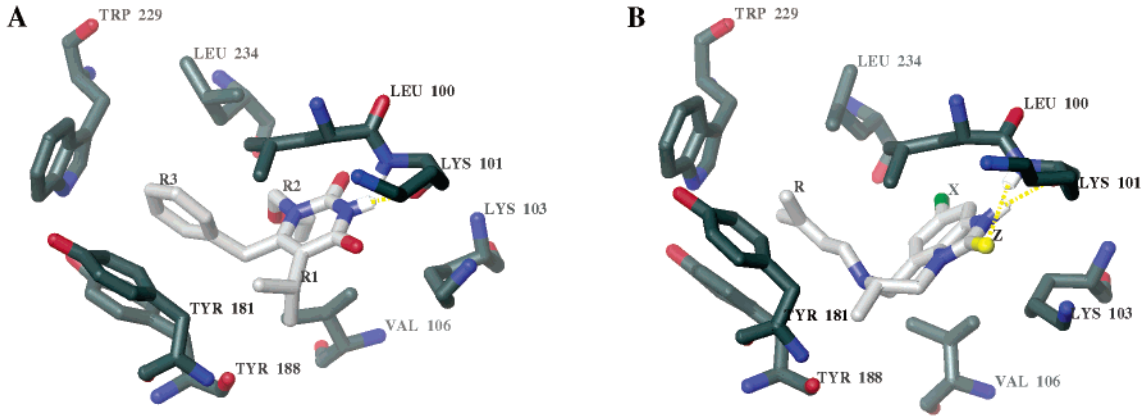

Figure 3. Representative conformations of the H11-RT (A) and 1a-RT (B) complexes extracted from the MD trajectories. Carbon atoms of the ligand are in light gray; those of the protein are in dark gray. Oxygen atoms are red, nitrogen atoms are blue, hydrogen atoms are white, the sulfur atom is yellow, and the chlorine atom is green. Hydrogen bonds are yellow.

HEPT compounds. However this trend is reversed when the estimator EC + 2ES, which takes into account desolvation, is considered, suggesting that, due to desolvation penalties, stronger ligand−receptor electrostatic interactions do not necessarily lead to better binding. Indeed the sign of the EC + 2ES estimator shows that generally this measure of the electrostatic component of the work of inserting the ligand into the receptor favors binding for the HEPT compounds (negative values), whereas it disfavors binding for the TIBO compounds (positive values). The EC + 2ES electrostatic estimator is a better predictor for the HEPT binding energies than for the TIBO binding energies. On the other hand, the EC + 2(ES − EL) estimator, which models the transfer of the ligand from solution into the receptor and also incorporates the loss of ligand−solvent interactions, is a better estimator for the binding energies of the TIBO compounds than the HEPT compounds. Finally the ligand hydration free energy estimator EL + VL + CL alone is rather weakly correlated to the HEPT and TIBO binding affinities.

These observations indicate that the energetic balance that drives binding is very different between these two sets of ligands and suggest that accurate modeling of both sets of ligands using a single regression equation would be difficult. It is also apparent that the HEPT compounds constitute a less challenging data set than the TIBO compounds. The HEPT binding free energies in general can be predicted fairly accurately using a single descriptor (the LJ descriptor), whereas for the TIBO compounds there is no single strong predictor of binding free energies. Successful prediction of the TIBO binding free energies must hinge on the interplay between the various predictors rather than on any single predictor.

Analysis of the calculated predictors for the TIBO compounds (Table 4) reveals a few ligands that stand out from the rest. Ligands 10g, 10h, and 15c have much more favorable estimated electrostatic hydration free energies (the EL descriptor) than all the other ligands. These unusually large ligand solvation free energies disfavor binding and would lead to the prediction of poor binding affinities if not counterbalanced by equally favorable ligand−receptor Coulomb interactions (the EC descriptor) and residual complex−solvent interactions (the ES descriptor). The 10g and 10h ligands are characterized by less positive ES values than the other ligands, but the EC descriptors are of equal magnitude as the other ligands. This results in unusually unfavorable electrostatic contributions to binding (measured for example by the EC + 2(ES − EL) estimator) which are inconsistent with the measured relatively large potency of these ligands. Indeed these two ligands show up as outliers in all of the LIE models we have tried. Both of these ligands have the iodo-substitution in the X position. This indicates a possible deficiency of the OPLS-AA/AGBNP force field for iodo-substitutions. Hence, suspecting inaccurate energetic modeling, we removed ligands 10g and 10h from the LIE training set. Similar observations apply to ligand 15c except that for this ligand the unfavorable electrostatic contribution to binding is consistent with its poor measured inhibition activity. Nevertheless this ligand appears as a clear outlier in some of the LIE models we investigated. The large predicted electrostatic hydration free energies of 15c is related to the presence of the acetoamido group in the X position. The acetoamido group, due to its large dipole moment, is by far the most polar substituent in this class of ligands. It is likely that the binding mode of 15c differs from the binding mode of 1a on which the structural model for the HIV-RT/15c complex is based. For this reason we have decided to remove the 15c ligand from the training set as well. The final TIBO training set consists of 37 ligands.

5.4. LIE Regression Models. The descriptors of eqs 33−40 are combined based on eqs 31 and 32 to construct two classes of models. The first class, based on the assumption that the same linear response proportionality coefficients apply to the work of inserting the ligand in the receptor and in water, is described by the general regression equation

$$\Delta F_b = \beta[\mathrm{EC}] + \beta'[\mathrm{ES} - \mathrm{EL}] + \alpha[\mathrm{LJ} + \mathrm{V} - \mathrm{VL}] + \gamma[\mathrm{C} - \mathrm{CL}] + \delta \quad (41)$$

where ΔF_b is the binding free energy, and the estimators result from the combination of descriptors of the complex (EC, ES, LJ, V, and C) and descriptors of the free ligand (EL, VL, and CL).

The second class of LIE models, given by the general expression

$$\Delta F_b = \beta[\mathrm{EC}] + \beta'[\mathrm{ES}] + \alpha[\mathrm{LJ} + \mathrm{V}] + \gamma[\mathrm{C}] - \omega[\mathrm{EL} + \mathrm{VL} + \mathrm{CL}] + \delta \quad (42)$$

treat the process of inserting the ligand in solution independently from the work of inserting the ligand in the receptor pocket. In this class one of the LIE estimators is the hydration free energy of the ligand, EL + VL + CL, estimated by the implicit solvent model with the LIE proportionality coefficient ω. All the other estimators include descriptors related only to the complex.

In addition to the electrostatic, van der Waals, and cavity LIE proportionality coefficients, the LIE regression equations we have investigated include an adjustable intercept parameter δ. Inclusion of this parameter is, in our view, crucial to the success of LIE paramterizations. The intercept parameter absorbs effects that influence equally absolute binding free energies and are therefore inconsequential for the ultimate goal of estimating relative binding free energies of a family of similar ligands. It is unreasonable and unnecessary to assume that linear response is applicable to the full process of ligand formation in the receptor and in solution. The success of a LIE model which includes an intercept parameter hinges only on the applicability of linear response for the simpler process of mutating one ligand into another. The inclusion of the intercept parameter allows the LIE coefficients to reflect relative binding free energies rather than absolute binding free energies, leading to better accuracy in ligand ranking as well as values of LIE coefficients in better agreement with linear response predictions.

We have investigated several models based on eqs 41 and 42, including those with the maximum number of adjustable parameters allowed by the general expressions (5 for eq 41 and 6 for eq 42). We present results for the following LIE regression models with 3 or fewer adjustable parameters other than the intercept, one from the first class (eq 41) and two from the second class (eq 42):

$$\mathrm{MDL1}: \Delta F_b = \beta[\mathrm{EC} + 2(\mathrm{ES} - \mathrm{EL})] + \alpha[\mathrm{LJ} + \mathrm{V} - \mathrm{VL}] + \gamma[\mathrm{C} - \mathrm{CL}] + \delta \quad (43)$$

$$\mathrm{MDL2}: \Delta F_b = \beta[\mathrm{EC} + 2\mathrm{ES}] + \alpha[\mathrm{LJ} + \mathrm{V}] + \gamma[\mathrm{C}] - \frac{1}{2}[\mathrm{EL} + \mathrm{VL} + \mathrm{CL}] + \delta \quad (44)$$

$$\mathrm{MDL3}: \Delta F_b = \frac{1}{2}\left\{\frac{1}{2}[\mathrm{EC}] + [\mathrm{ES}] + [\mathrm{LJ} + \mathrm{V}] - [\mathrm{EL} + \mathrm{VL} + \mathrm{CL}]\right\} + \gamma[\mathrm{C}] + \delta \quad (45)$$

We have chosen these as the most representative models because the first, MDL1, is comparable to standard models used in previous LIE studies[3,4] and the second, MDL2, is based on the novel LIE formulation (eq 42) with the same number of adjustable parameters as MDL1. The third model, MDL3, is a minimally parametrized version of MDL2. MDL1 is based on eq 41 with $\beta' = 2\beta$, MDL2 is based on eq 42 with $\beta' = 2\beta$ and $\omega = 1/2$, and MDL3 is the same as MDL2 with the additional constraints that $\beta = 1/2$ and $\alpha = 1$. Further motivation for these choices of parameters is provided below and in the Discussion section, where we also discuss some of the results obtained with the other models we tested.

Table 5 shows the results of the LIE regressions for the HEPT and TIBO ligand sets using the models presented above. This table contains the results of fitting MDL1, MDL2, and MDL3 to the HEPT set, the TIBO set, and the two sets combined. It reports the values of the LIE coefficients obtained from multivariate linear least-squares fitting, the square of the correlation coefficient (R^2), and root-mean-square deviation (RMSD) between experimental and estimated LIE binding free energies. R_{pred}^2 and $RMSD_{pred}$ are the corresponding quantities for the jack-knife regression tests. In these tests a LIE model is fit to the ligand set removing each ligand in turn and then using the resulting LIE model to predict the binding free energy of the ligand that was left out. The jack-knife results give a better unbiased representation of the predictive ability and transferability of each model.

MDL1 closely follows corresponding formulas for LIE estimates of ligand binding from explicit solvent simulations. The electrostatic estimator (EC + 2(ES − EL)) is the difference between the bound and solution phases of the average electrostatic interaction energy between the ligand and the environment. In the bound form the electrostatic interaction energy is estimated as the sum of the ligand−receptor average electrostatic interaction energy plus twice the Generalized Born energy of the ligand in the complex as measured by the ES descriptor (eq 34). In the solution phase the average ligand-environment electrostatic energy is measured as twice the EL descriptor (eq 35). As discussed in the previous section and in the Appendix, the assumption that $\beta' = 2\beta$ in eq 41 puts the estimators corresponding to interactions between explicit atoms and the implicit solvent components on equal footing. To validate this choice we have tested two additional models based on eq 41, one setting $\beta' = \beta$ and another one in which the β and β' coefficients are allowed to vary independently. The results of these tests are discussed in the next section.

The second model (MDL2) has the same number of adjustable parameters as MDL1. It is based on eq 42 with $\beta' = 2\beta$ and the observation that setting $\omega = 1/2$ leads to more accurate results than choosing the value $\omega = 1$ suggested based on linear response theory (see Theory and Methods section). Indeed, as further discussed later, we find that, while their ratios agree with linear response predictions, the absolute values of the fitted β, β', and α LIE coefficients are approximately half their theoretical values. The final model (MDL3) exploits this observation by using only one adjustable coefficient (beside the intercept parameter): the surface tension parameter γ for creating the ligand cavity in the receptor. All the other parameters are set to half the values expected from linear response theory.

6. Discussion

6.1. Model Performance. MDL1 uses estimators that are designed to mimic the corresponding estimators commonly used in LIE explicit solvent studies. This model is also comparable to the implicit solvent LIE formulation introduced by Zhou et al.[4] The R^2 and RMSD values obtained in this work (see Table 5) for the HEPT compounds using MDL1 are 0.87 and 0.80 kcal/mol, respectively, compared

Table 5: Fitting Results for MDL1, MDL2, and MDL3 to the HEPT and TIBO Sets, Individually and Combined[b]

model	n_p	β	α	γ	δ	R^2	R_{pred}^2	RMSD	$RMSD_{pred}$
				HEPT Set (20 Ligands)					
MDL1	4	0.22	0.40	0.20	1.56	0.87	0.81	0.80	0.97
MDL2	4	0.21	0.45	−0.05	6.28	0.91	0.87	0.66	0.80
MDL3	2	1/4[a]	1/2[a]	−0.18	7.73	0.90	0.88	0.71	0.77
				TIBO Set (37 Ligands)					
MDL1	4	0.27	0.40	0.36	3.35	0.70	0.64	1.12	1.23
MDL2	4	0.23	0.45	0.19	5.80	0.73	0.66	1.07	1.19
MDL3	2	1/4[a]	1/2[a]	0.19	8.09	0.73	0.68	1.08	1.16
				HEPT+TIBO Set (57 Ligands)					
MDL1	4	0.15	0.11	0.29	0.21	0.55	0.48	1.43	1.54
MDL2	4	0.07	0.39	0.17	4.15	0.67	0.62	1.23	1.31
MDL3	2	1/4[a]	1/2[a]	0.004	7.34	0.07	−0.02	2.06	2.16

[a] Set value not allowed to vary during the fitting. [b] n_p is the number of adjustable parameters (including the intercept parameter), and β, α, γ, and δ are the LIE adjustable coefficients. R^2 is the square of the correlation coefficient of the fit, and RMSD is the root-mean-square deviation between the LIE equation and the experimental binding free energies. R_{pred}^2 and $RMSD_{pred}$ are the corresponding quantities for the jack-knife validations.

to $R^2 = 0.78$ and RMSD = 1.07 kcal/mol reported by Zhou et al. for the same set of ligands. Zhou et al. report that the binding free energies of MKC-442 (compound H11) and H14 were significantly underpredicted by their models. These two ligands are correctly predicted by the MDL1 model (the deviation between predicted and experimental binding free energies for these two compounds are 0.58 and 0.03 kcal/mol, respectively). The superior performance of the MDL1 LIE model for the HEPT compounds relative to the LIE results reported by Zhou et al. is likely due in part to the differences in the expressions for the LIE estimators but may also reflect differences in the conformational ensemble sampled by the two studies. Our MD simulations, including thermalization and equilibration phases, are longer than those performed by Zhou et al., and the implicit solvent models employed in the two studies are different: SGB/SA[61] in the study of Zhou et al. and AGBNP[35] in the present study. The electrostatic LIE fitting coefficients β we obtained with MDL1 (0.22) for the HEPT compounds is in reasonable agreement with the value obtained by Zhou et al., but the van der Waals and cavity parameters α and γ are significantly different. The smaller values for the α parameter obtained by Zhou et al. are most likely due to differences in the corresponding implicit solvent functional form. The large differences in LIE γ fitting coefficients are mainly due to the differences in the values of the cavity surface tension employed in the two studies.

The same set of HEPT compounds were also studied previously using extended LIE models based on descriptors obtained from explicit solvent simulations.[19,53] Based on 3 descriptors (variation in number of hydrogen bonds upon binding, ligand−protein Lennard-Jones interaction energy, and variation in exposed hydrophobic surface area), and excluding H17, the reported correlation coefficient of the best performing extended LIE model[19] is $R^2 = 0.85$ and RMSD = 0.87 kcal/mol similar to our MDL1 results ($R^2 = 0.87$, RMSD = 0.80 kcal/mol).

Subsets of the TIBO compounds studied here have been previously studied by LIE modeling. Smith et al.[55] studied 12 TIBO compounds using explicit solvent LIE models similar to our MDL1 model. Smith et al. obtain RMSDs between 0.9 and 1.0 kcal/mol. Rizzo et al.[19] considered 22 TIBO compounds using two descriptors (variation in number of hydrogen bonds upon binding and ligand−protein Lennard-Jones interaction energy), obtaining, after excluding two outliers, $R^2 = 0.79$ and RMSD = 0.75 kcal/mol. In comparison, results for the same set of ligands using MDL1 are $R^2 = 0.88$ and RMSD = 0.71 kcal/mol. The TIBO set studied here (37 compounds) is larger and more diverse than in the two previous LIE studies; nevertheless the MDL1 LIE model achieves good accuracy for these ligands (see Table 5).

The results of fitting MDL1, MDL2, and MDL3 (eqs 43−45) to the HEPT and TIBO compounds separately and combined are listed in Table 5. MDL3 (see also Figure 2) yields the highest prediction accuracy for the HEPT and TIBO compounds separately. This model is based on eq 42 with only two adjustable parameters (γ and the intercept δ); all the other LIE coefficients are set to half the linear response theory predictions ($\beta = 1/4$, $\beta' = 1/2$, $\alpha = 1/2$, $\omega = 1/2$). R_{pred}^2 of MDL3 for the HEPT compounds is 0.88 and $RMSD_{pred}$ is 0.77 kcal/mol which are the best among all the models tested, including those with up to 6 adjustable parameters. The corresponding values for fitting MDL3 to the TIBO compounds are $R_{pred}^2 = 0.68$ and $RMSD_{pred} = 1.16$ kcal/mol, again the best among all the models tested. This result highlights the benefit of minimizing the number of adjustable parameters in LIE models. Increasing the number of parameters to improve the accuracy of the fit can affect the transferability of the model even among closely related ligands, such as those of the HEPT set.

Based on the jack-knife indicators, the best model for the combined HEPT+TIBO set is MDL2 ($R_{pred}^2 = 0.62$ and $RMSD_{pred} = 1.31$ kcal/mol). This model is slightly superior to the LIE model with 6 adjustable parameters we tested based on eq 42, two more than MDL2, and significantly better than MDL1. MDL3 fails completely in reproducing the binding free energies of the combined HEPT+TIBO set, as expected by the very different values of the γ LIE coefficient obtained by fitting MDL3 to the two ligand sets separately. This issue is discussed further in the next section. The values of the LIE coefficients of MDL3 for the combined

HEPT+TIBO set are similar to those of MDL2 except for β. It appears that MDL2 is able to fit the binding free energies of the combined set by employing a smaller electrostatic LIE coefficient than MDL3.

In general the models that allow for differences in the response of the receptor and water media (MDL2 and MDL3) perform better than MDL1 that does not. For example, for the HEPT compounds the RMSD of MDL2 is 0.66 kcal/mol as compared to 0.80 kcal/mol of MDL1 which has the same number of parameters. The superior performance of MDL2 over MDL1 is particularly noticeable for the combined HEPT+TIBO set. The model with the highest prediction accuracy for the HEPT and TIBO set taken separately is MDL3 which is based on the decoupling of the solution and receptor environment insertion processes. This is a further indication that the decoupling of the receptor and solution insertion processes can lead to LIE models with superior prediction accuracy.

Consistent with the linear response prediction that the electrostatic implicit solvent descriptor should be weighted twice the receptor−ligand Coulomb interaction energy, models we tested with $\beta' = \beta$ in eqs 41 and 42 perform significantly worse than MDL1 and MDL2 for both the HEPT and TIBO compounds. This is further confirmed by the results of models in which the β and β' LIE coefficients are optimized independently. In these models the optimal value of β was found to be roughly half the value of β'.

In principle, because the binding free energy is the sum of the work of inserting the ligand in the protein environment and the work of inserting the ligand in water (the hydration free energy), the LIE coefficient, ω, associated with the estimated hydration free energy should be set to 1. However we find that the accuracy of models that set $\omega = 1$ is rather poor. Much better results are obtained, as in MDL2, when ω is set to half the theoretical value. Indeed it is generally observed that, when the LIE coefficients in eq 42 are allowed to vary, they assume optimal values approximately half their theoretical values; this issue is discussed below. Two models we tested based on eq 42 which allow for optimization of ω and both ω and β' have resulted in modest gains in term of accuracy over MDL2 for which ω is fixed at one-half its theoretical value.

The TIBO set studied here is larger (37 ligands) and more diverse than previous LIE studies. The LIE models for the TIBO class of NNRTI's presented here are, to our knowledge, the best reported in terms of prediction accuracy, as measured by the jack-knife and R^2 and RMSD values. The models have few outliers and reproduce well the main trends in this class of compounds. The generally superior binding affinity of compounds with sulfur at the Z position relative to oxygen is reproduced. Analysis of the predictors of the pairs of TIBO compounds that differ only on the substitution at the Z position reveals that the C predictor (the average cavity hydration free energy of the complex relative to the that of the receptor) is mainly responsible for the ability of the LIE models to distinguish the sulfur and oxygen substituents. This points to hydrophobicity as the main cause of the stronger binding affinity of sulfur-containing TIBO inhibitors; whereas the sulfur atom at the Z position forms electrostatic and hydrogen bonding interactions with the receptor of equal magnitude as oxygen in the same position, the sulfur atom in this position is capable of burying more solvent-exposed receptor surface area. The models also reproduce the higher binding affinity obtained with the 8-Cl substitution relative to the 9-Cl substitution. We ascribe this result to the interaction between the chlorine atom at the 8 position with PHE227 also seen in crystal structures.[57] This contact causes a shift of the position of the ligand leading to increased contacts between the hydrophobic side chain of the ligand and the hydrophobic pocket formed by TYR181, TYR188, and TRP229.

MDL2 is able to fit well the combined HEPT and TIBO sets (57 ligands). Particularly encouraging are the results with MDL3. This model, based on only one adjustable parameter (excluding the intercept), has the highest prediction accuracy for the HEPT and TIBO compounds separately.

Our results suggest that the LIE formalism described here using the OPLS-AA/AGBNP effective potential improves on the results we obtained previously using the standard LIE formalism and the SGB/NP implicit solvent potential[4] and is competitive with corresponding results in explicit solvent.[19,53,55] Although differences in ligand force field parametrizations could exist, the major difference between this and previous LIE studies of HEPT and TIBO HIV-RT inhibitors is the modeling of the solvent. The AGBNP implicit solvent model does not provide a description of hydration forces with the same detail as explicit solvent models. Nevertheless molecular dynamics sampling with implicit solvent is able to explore a more diverse ensemble of conformations than a comparable calculation of the same CPU cost using an explicit solvent representation. A more complete representation of the ensemble of conformations of the ligand and the complex may be the basis for the superior results we obtained relative to similar explicit solvent studies.[19,55]

6.2. Physical Interpretation of LIE Models. The values of the β and α parameters for all models tested and the values of γ for MDL1 are positive. This implies that these models, in accordance with physical intuition, predict that ligands that have stronger interactions with the receptor and smaller desolvation penalties tend to be better binders. The seemingly counterintuitive negative value of γ obtained with MDL2 and MDL3 for the HEPT compounds is discussed below.

The LIE coefficients for MDL1 correspond to the difference between the work of inserting the ligand in the receptor and in water assuming that the same linear response proportionality coefficient applies to both processes. In the MDL2 and MDL3 models, however, the process of insertion of the ligand in the receptor is decoupled from the process of insertion in water, thus allowing for differences in the effective linear response of the receptor and water media. We see from Table 5 that the value of β differs little from one class of models to the other, suggesting that the receptor environment responds linearly to the ligand electrostatic field in a manner similar to water. A similar behavior can be observed for the α coefficient which measures the response of the water and receptor environment to the introduction of ligand−solvent and ligand−receptor van der Waals

interactions. Despite these similarities, as discussed above, the models that allow for differences in the response of the protein receptor and water (MDL2, MDL3, and related models) generally perform better.

In models such as MDL1 based on eq 41 the γ coefficient corresponds to the C−CL estimator, whereas in models based on eq 42, such as MDL2 and MDL3, it corresponds to the C descriptor, the average of the difference of the cavity hydration free energy of the complex with and without the ligand. The C descriptor approximately measures the free energy gain of burying the receptor residues lining the binding site. Thus, we would expect a positive linear correlation ($\gamma > 0$) between the C descriptor, which is always negative (see Tables 3 and 4), and the binding free energy. This is indeed the case for MDL2 and MDL3 for the TIBO compounds. However the same models applied to the HEPT compounds yield a negative γ coefficient (see Table 5). For both ligand sets the value of γ for MDL2 and MDL3 is significantly smaller than for MDL1. It appears therefore that in these models the γ coefficient absorbs physical effects correlated to the size of the binding site which oppose binding and are not directly related to the hydrophobic effect.

A likely candidate in this role is the reorganization free energy of the receptor, that is the work required to deform the binding site region to accommodate the ligand. It is reasonable to assume that the larger the binding site the larger the work required to deform the receptor structure. This contribution effectively reduces the benefit of having a large ligand−receptor contact surface area and could result in the smaller values for the γ LIE coefficients obtained for MDL2 and MDL3 as compared to MDL1, which adopts estimators that combine solution and receptor environments descriptors and is therefore less able to capture these effects. We hypothesize that for the HEPT compounds the conformational strain contribution overcomes the favorable hydrophobic desolvation effect, resulting in a negative value of γ. The prediction that the receptor reorganization free energy plays a more important role for the HEPT compounds than the TIBO compounds could also be a consequence of the wider distribution of ligand sizes of the HEPT compounds as measured by the ligand cavity formation descriptor which is proportional to the solute surface area. The range of variation of the CL descriptor of the HEPT compounds is 13.0 kcal/mol as compared 5.8 kcal/mol of the TIBO compounds (see Tables 3 and 4). The larger variation of receptor reorganization free energies embedded in the experimental binding free energies of the HEPT compounds makes the statistical fit of this term via linear regression more significant.

Although deviations from ideality of LIE parameters can also be caused by the residual average electrostatic potential created by the receptor in absence of solute charges,[42] we do not believe that this is the case in the present system. As eq 12 shows, linear response theory predicts that a nonzero residual electrostatic interaction energy ($\langle V \rangle_0$ in eq 12) tends to increase the value of the corresponding electrostatic LIE coefficient from its ideal value of 1/2. If the electrostatic LIE descriptor EC+2EL is negative (as for most HEPT compounds), this effect would appear, to an LIE model that adopts an ideal LIE electrostatic coefficient of 1/2, as an unaccounted term *favorable* to binding. It seems therefore unlikely that the residual electrostatic field of the receptor is the origin of the observed negative value of γ resulting from an unaccounted effect *unfavorable* to binding, such as the reorganization free energy of the receptor.

The results obtained with LIE models that do not include the factor of 2 for the implicit solvent electrostatic estimators are significantly inferior than the other models with the same number of adjustable parameters. It is predicted based on linear response theory that β should be half the value of β'. This is indeed approximately the case for models in which β (which corresponds to the ligand−receptor electrostatic interaction energy) is allowed to vary independently from β' (the coefficient corresponding to the implicit solvent electrostatic estimator). Contrary to linear response predictions however β is always smaller than the theoretical value of 1/2, and β' is always smaller than the theoretical value of 1. This is the case for the calculated α and ω parameters as well. It appears that the calculated parameters differ from their theoretical values by a constant proportionality factor. This would occur if a linear relationship with a proportionality coefficient different from 1 exists between the effective binding free energies calculated from the IC_{50}'s as $kT \ln IC_{50}$ and the actual binding free energies. We plan to further investigate this issue by studying complexes for which direct measurements of the binding free energies have been reported. The IC_{50}'s used in this work have been obtained from cell survival assays,[53−55] rather than enzymatic rate inhibition assays which are more directly related to the binding affinity. In cell-based assays the IC_{50} is calculated from the number of cells in a colony that remain viable after infection with the HIV-1 virus in the presence of the inhibitor. A number of environmental factors such as cell absorption and metabolism and molecular factors[62] such as stoichiometry of binding and induced RT dimerization could affect the observed effective binding free energies. The results for HIV-1 RT NNRTI systems similar to those studied here[5,55] are in general agreement with our finding that, even though the optimal values of β and α are smaller than expected, their relative magnitude is consistent with linear response predictions.

To support the hypothesis that the binding free energies calculated from the measured IC_{50}'s are proportional to the actual binding free energies, we show in Table 6 the ratios between the β, β', and ω parameters and the value of the α parameter obtained from a series of models (including those in Table 5). The theoretical values of these ratios based on linear response are $\beta/\alpha = 1/2$, $\beta'/\alpha = 1$, and $\omega/\alpha = 1$. It can be clearly seen from Table 6 that the relative magnitudes of the LIE coefficients are in good agreement with the theoretical predictions even though their absolute values do not. Although this result indicates that the effective free energies of binding derived from the IC_{50}'s deviate from the actual binding free energies by a constant proportionality factor, further investigations are needed to confirm this hypothesis.

The constant proportionality factor between the theoretical and calculated LIE coefficients appears to be close to 1/2.

Table 6: Ratios between the LIE Parameter β, β', and ω and the van der Waals LIE Parameter α Extracted from the Fits to the Experimental HEPT and TIBO Binding Data for the Model Listed Compared to Linear Response Theoretical Values

	HEPT set			TIBO set		
model	β/α	β'/α	ω/α	β/α	β'/α	ω/α
theory	0.50	1	1	0.50	1	1
MDL1	0.55			0.67		
MDL1a[a]	0.79	1.18		0.67	1.37	
MDL2	0.47		1.11	0.51		1.11
MDL2a[b]	0.45		1.07	0.55		1.24
MDL2b[c]	0.68	1.06	1.18	0.64	1.18	1.24

[a] Eq 41. [b] Eq 42 with $\beta' = 2\beta$. [c] Eq 42.

The values of the α, β', and ω coefficients are consistently near 0.5 (see Table 5), whereas the theoretical value for these parameters based on linear response is 1. Models in which ω is set to 1 perform significantly worse (data not shown) than MDL2 in which ω is set to 1/2 (Table 5). Similarly, the calculated value of the β coefficients is close to 1/4 as compared to the theoretical value of 1/2. Based on these observations we have attempted to construct a minimally parameterized LIE model for both the HEPT and TIBO compounds by setting the LIE coefficients to their theoretical linear response values reduced by a factor of 1/2. The remaining two parameters for which we do not have a firm theoretical prediction (the parameter γ corresponding to the cavity descriptor C, and the intercept δ) were allowed to vary to fit to the experimental binding free energies. This model (MDL3 in Table 5) is the most successful in predicting the binding free energies of the HEPT and TIBO compounds separately as measured by the jack-knife correlation coefficient R_{pred}^2 and root-mean-square deviation $RMSD_{pred}$.

A noticeable difference between the LIE coefficients of model MDL3 for the HEPT and TIBO compounds is the value of γ, which is negative ($\gamma = -0.18$) for the HEPT compounds and positive ($\gamma = 0.19$) for the TIBO compounds. The high statistical significance of this difference is highlighted by the poor results obtained when fitting MDL3 to the combined HEPT and TIBO set. In this fit the γ coefficient is close to zero, an intermediate value between the value appropriate for the HEPT compounds and that appropriate for the TIBO compounds, and, as a result, the model is unable to fit accurately the experimental binding free energies. As discussed above we believe that this difference is due to the reorganization free energy of the receptor which opposes binding of the HEPT analogues more than the TIBO analogues, whereas hydrophobicity favors association relatively equally. Based on this result we predict that a hypothetical increase of ligand size, which leaves electrostatic and van der Waals receptor−ligand interactions unchanged, would have opposite effects on the binding affinities of the HEPT and TIBO analogues. For the TIBO analogues the hydrophobic gain due to the increase in the amount of receptor surface area buried by a larger ligand overcomes the opposing effect due the increase of receptor reorganization free energy, leading to stronger binding. For the HEPT analogues, instead, the hydrophobic gain is more than offset by the increase of receptor reorganization free energy, leading to weaker binding. In practice, however, because it is not possible to modify the size of the ligand without also affecting electrostatic and van der Waals receptor−ligand interaction energies, all of the energetic components of the LIE model need to be considered to accurately predict the effect of ligand modifications.

The success we obtained with MDL3 in predicting the binding free energies of the HEPT and TIBO compounds based on theoretically derived parameters and only one adjustable LIE coefficient (excluding the intercept) raises the question of whether this result is a direct consequence of the receptor and solution environments obeying linear response or simply a coincidental occurrence for the receptor system and ligand sets we analyzed. The results of previous LIE studies on a variety of ligand binding systems are mainly inconclusive on this issue, partly due to difficulty of interpreting results obtained with different LIE regression expressions.

A LIE explicit solvent study of P450cam complexes by Paulsen and Ornstein[63] has shown that values of LIE coefficients near to their theoretical values of $\alpha = 1$ and $\beta = 1/2$ reproduced well the experimental binding free energies, whereas in a related study Almlöf et al.[44] reproduced the binding free energies of a similar set of inhibitors of P450cam with a significantly smaller value of the α LIE coefficient. As noted by Amlöf et al.[44] the difference between the values of α between the two studies is due to the intercept parameter, considered only in the latter study, whose optimal value was found to depend on the hydrophobicity of the receptor site.[44] Wang et al.[64] have investigated different values of the van der Waals parameter α in conjunction with β set to 1/2 without and intercept parameter. They found that for various ligand−protein complexes the optimal value of α varied from 1 to e 0.1 depending on the hydrophobicity of the binding pocket. In a related study Wang et al.[65] have compared explicit solvent LIE predictions to rigorous free energy thermodynamic integration (TI) calculations for complexes with streptavidin and established that for these systems values of α and β near their theoretical values reproduced well the experimental binding free energies for ligands for which LIE and TI produce consistent results. Cases in which TI produced results in better agreement with the experiments could be rationalized by the inadequacy of the LIE model to properly take into account the free energy cost for reorganizing the receptor pocket.[65] Earlier LIE studies[2,9] reported smaller optimal values for the α LIE parameter when setting β to 1/2. Clearly, much remains to be done to better understand the factors that influence the values of the LIE coefficients obtained on different systems with different LIE regression equations.

It is conceivable that LIE regression models, such as those adopted in the present work, which include an estimator for the work of cavity formation[3,4] are more likely to yield electrostatic and van der Waals LIE coefficients consistent with linear response predictions. The values of electrostatic and especially van der Waals LIE coefficients obtained using LIE models without an explicit cavity formation estimator,[44] and especially those without an intercept parameter,[63,65] are

instead more likely to absorb effects that are not directly related to electrostatic and van der Waals interactions and are thus not as amenable to rationalization and generalization in terms of linear response theory. Furthermore, because solute−solvent van der Waals interaction energies are not perfectly correlated with the solute surface area,[23,66,67] surface area-based cavity estimators provide nonredundant information content in addition to the van der Waals estimator, potentially leading to better descriptive accuracy and transferability. However, although it has been generally recognized that the free energy of cavity formation in water approximately scales as the surface area of the solute,[23,38] the validity of using a surface area estimator to describe the work of ligand cavity formation in the receptor pocket remains to be fully addressed.

7. Conclusions

In this paper we have addressed a number of outstanding issues in regard to the theory and practice of LIE modeling for binding free energy prediction. We have reviewed and clarified the linear response theory on which LIE methods are based. Following linear response formalism, we derived expressions for the LIE estimators when the solvent is treated implicitly. We showed that these estimators include descriptors related to the desolvation of receptor atoms which were not considered in a previously reported implicit solvent LIE model.[4] The form of the estimators we derived are consistent with those proposed previously in the context of the LRA method and the PDLD implicit solvent models[10] and for the LIE electrostatic estimator with a GB model.[12] We have also developed a novel class of LIE models that, contrary to current practice, attempt to model independently the processes of insertion of the ligand in the receptor and in solution. These models are motivated by the fact that, potentially, the receptor and the solvent respond differently to the introduction of the ligand.

We have applied these ideas to the problem of the binding free energy estimation of a series of NNRTI inhibitors of HIV-1 RT. LIE descriptors were collected from 57 molecular dynamics simulations of HIV-1 RT complexed with 20 HEPT inhibitors and 37 TIBO inhibitors. Based on the measured binding affinities and the calculated descriptors we developed a series of LIE models. We presented results for three of these models and tested several others. The first model, MDL1, is comparable to previously reported LIE studies for NNRTI binding in both explicit and implicit solvent. Relative to these studies, the predictive accuracy of our MDL1 model is generally superior when applied to the same ligand sets, suggesting that the AGBNP implicit solvation model provides sufficient accuracy for LIE modeling. This result also indicates that the LIE estimators we derived are more appropriate to describe the energetics of the binding process in implicit solvent than previously reported alternatives. The second and third models, MDL2 and MDL3, are novel models designed to treat the hydration free energy of the ligand and the work of inserting the ligand in the receptor independently. MDL2, which has the same number of parameters as MDL1, is found to be superior to MDL1 and to LIE models developed by others for predicting the binding affinities of the HEPT and TIBO analogues. This result may indicate that LIE models that treat the two insertion processes independently can lead to better prediction accuracy. MDL3, a minimally parametrized version of MDL2 in which some parameters are set to their values predicted by linear response, is found to be superior to MDL1 and MDL2 in terms of predictive ability (jack-knife tests) for the HEPT and TIBO sets individually but not for the two sets combined. We hypothesize that this is caused by the larger sizes of some of the HEPT compounds which induce steric strain of the receptor; an effect which is not taken explicitly into account by the LIE models.

We examined the values of the LIE coefficients obtained from the regression analysis and established that, although they are smaller than expected, their relative magnitudes generally conform to linear response predictions. We are planning calculations to test the applicability of linear response to other protein−ligand binding systems by computing directly the relative free energies of inserting ligands in solution and the protein environment and comparing the results to the corresponding first order cumulant descriptor. The discovery that linear response behavior is generally applicable to protein−ligand binding could provide the basis for the development of minimally parametrized LIE models. Given the substantial reduction of adjustable parameters achievable when assuming linear response for some of the interaction energy contributions and the potential improvement in predictive ability, minimally parametrized models based on linear response, such as the ones presented here, offer a productive route to binding free energy predictions using sparse experimental binding assay data for lead optimization in structure-based drug design.

Acknowledgment. This work has been supported in part by a grant from the National Institutes of Health (GM30580). We thank Dr. Anthony Felts and Dr. Zhiyong Zhou for helpful discussions.

Appendix

In this Appendix we derive eq 30. Let us first consider the process of turning on the ligand charges in the receptor environment. Conceptually we will divide this process into two steps. First the electrostatic ligand−receptor interactions are turned on and then interactions between the ligand and the implicit solvent continuum are turned on. If we assume that the receptor responds to the perturbation as a perfect linear dielectric, the free energy change, $\Delta F_{el}^{(1)}$, corresponding to the first step is approximately

$$\Delta F_{el}^{(1)} \simeq \beta \langle V_{el} \rangle_1 \tag{46}$$

where $\langle V_{el} \rangle_1$ is the average of the ligand−receptor Coulomb interaction energy in the absence of ligand-continuum solvent electrostatic interactions, and $\beta = 1/2$ based on eq 13 under the present assumptions. Based on linear response the free energy corresponding to turning on the ligand−implicit solvent continuum electrostatic interactions is

$$\Delta F_{el}^{(2)} = \beta' \langle G_{el}^{c} - G_{el}^{c'} \rangle_c \tag{47}$$

where G_{el}^{c} is the Generalized Born energy of the receptor−ligand complex and $G_{el}^{c'}$ is the Generalized Born energy of complex when the ligand charges are turned off, and the average is taken in the ensemble in which the ligand is fully charged. To derive eq 47 we write the λ-dependent potential for this process as $U(\lambda) = U_0 + \lambda V$, where $U_0 = U_{expl} + G_{cav} + G_{vdw} + G_{el}^{c'}$ is the reference potential and $V = G_{el}^{c} - G_{el}^{c'}$ is the perturbation. When the ligand and the receptor are both assumed rigid, the free energy change for adding electrostatic solute-continuum solvent interactions reduces to $\Delta F^{(2)}_{el} = G_{el}^{c} - G_{el}^{c'}$, which is the same as eq 47 with $\beta' = 1$ and with the average replaced by the constant value of the argument. Hence, based on the discussion presented above, in the general case when the ligand and the receptor are flexible and linear response applies, we expect the optimal value of the β' coefficient in eq 47 to assume values near 1.

Thus, assuming that in eq 46 the mean ligand−receptor electrostatic interaction energy $\langle V_{el} \rangle$ is the same whether the ligand-continuum solvent electrostatic interactions are present or not and assuming that $\beta' = 2\beta$, eqs 46 and 47 can be combined to give the linear response expression for the free energy of turning on the ligand charges in the receptor environment

$$\Delta F_{el} \simeq \beta \langle V_{el} + 2(G_{el}^{c} - G_{el}^{c'}) \rangle_{c} \quad (48)$$

where the average is taken with the ligand fully interacting with the receptor and solution environments, and c′ corresponds to the state in which the ligand is uncharged. It is straightforward to show using eq 18 that the difference between G_{el}^{c} and $G_{el}^{c'}$ for a given ligand−receptor complex conformation is the sum of the GB self-energies of the ligand atoms plus the sum of the GB pair energies between ligand atoms and between ligand atoms and receptor atoms.

The expression for the LIE estimator for the free energy of adding the van der Waals ligand−receptor and ligand−water interactions to the ligand cavity involves the attractive part of the Lennard-Jones potential and the implicit solvent solute−solvent van der Waals interaction energy G_{vdw}. Following the same derivation as for the electrostatic case and assuming $\alpha \simeq 1$ for both the explicit and implicit contributions, we obtain

$$\Delta F_{vdw} = \alpha \langle V_{LJ}^{vdw} + (G_{vdw}^{c'} - G_{vdw}^{c''}) \rangle_{c'} \quad (49)$$

where V_{LJ}^{vdw} is the sum of the attractive components of the Lennard-Jones interactions[68] between ligand and receptor atoms, $G_{vdw}^{c'}$ is the solute−solvent van der Waals interaction energy of the receptor−ligand complex, $G_{vdw}^{c''}$ is the solute−solvent van der Waals interaction energy of the complex in the absence of van der Waals interactions between the ligand and the solvent, and the average is taken in the state c′ with the uncharged ligand and with full ligand−receptor and ligand−solvent van der Waals interactions.

We now consider the work of creating the ligand cavity within the receptor site. According to the scheme developed above we should consider as an LIE estimator for this process the average of the difference between the solvent potential of mean force in the presence of the solute cavity and in the absence of the solute cavity plus repulsive interactions between the ligand atoms and receptor atoms. As above the average is assumed to be taken over the ensemble of conformations generated when the ligand cavity is present. The solvent potential of mean force difference includes two components: the change in G_{cav} in going from the receptor without the ligand cavity and the receptor with the ligand cavity, and, in addition, the change of the receptor−solvent van der Waals interaction energy and the change of Generalized Born energy of the receptor caused by the increase of the receptor atoms' Born radii due to the introduction of the ligand cavity. The explicit ligand cavity−receptor interactions are modeled using the repulsive component of the Lennard-Jones potential according to the WCA decomposition.[68] Finally, assuming that linear response applies to processes involving V_{LJ}^{rep} and the van der Waals and electrostatic implicit solvent components with the same proportionality coefficients as in eqs 48 and 49, we obtain the following expression for introducing the ligand cavity in the receptor

$$\Delta F_{cav} = \alpha \langle V_{LJ}^{rep} + (G_{vdw}^{c''} - G_{vdw}^{p}) \rangle_{c''} + \beta \langle 2(G_{el}^{c''} - G_{el}^{p}) \rangle_{c''} + \gamma \langle G_{cav}^{c''} - G_{cav}^{p} \rangle_{c''} \quad (50)$$

where V_{LJ}^{rep} is the sum of repulsive Lennard-Jones interactions between the ligand and receptor atoms, $G_{vdw}^{c''}$ and $G_{el}^{c''}$ are, respectively, the solute−solvent van der Waals interaction energy and the Generalized Born energy of the complex with the ligand cavity with the ligand-environment van der Waals and electrostatic interactions turned off. G_{vdw}^{p} and G_{el}^{p} are, respectively, the solute−solvent van der Waals interaction energy and the Generalized Born energy of the receptor conformation obtained from the conformation of the complex after removal of the ligand, $G_{cav}^{c''}$ is the cavity free energy of the receptor−ligand complex, G_{cav}^{p} is the cavity free energy of the complex after removal of the ligand, and $\langle \cdots \rangle_{c''}$ indicates averaging over complex conformations with the ligand cavity.

When considering the three estimators from eqs 48−50 we now make the approximation to evaluate all of the averages in the final state in which the ligand fully interacts with the environment (state *c*). We choose to combine the repulsive and attractive WCA components, V_{LJ}^{rep} and V_{LJ}^{vdw}, of the receptor−ligand Lennard-Jones interaction energies to form the total ligand−receptor Lennard-Jones interaction energy, V_{LJ}. Also, when combining the cavity and electrostatic descriptors the $G_{vdw}^{c''}$ and $G_{el}^{c''}$ terms cancel out. We then obtain the following expression for the LIE regression equation for the free energy for creating the ligand in the receptor site

$$\Delta F_{c} \simeq \alpha \langle V_{LJ} + G_{vdw}^{c} - G_{vdw}^{p} \rangle_{c} + \beta \langle V_{el} + 2(G_{el}^{c} - G_{el}^{p}) \rangle_{c} + \gamma \langle G_{cav}^{c} - G_{cav}^{p} \rangle_{c} \quad (51)$$

which is eq 30.

References

(1) *Free Energy Calculations in Rational Drug Design;* Reddy, M. R., Erion, M. D., Eds.; Springer-Verlag: 2001.

(2) Åqvist, J.; Medina, C.; Samuelsson, J.-E. *Protein Eng.* **1994**, *7*, 385−391.

(3) Jones-Hertzog, D. K.; Jorgensen, W. L. *J. Med. Chem.* **1997**, *40*, 1539−1549.

(4) Zhou, R.; Friesner, R. A.; Ghosh, A.; Rizzo, R. C.; Jorgensen, W. L.; Levy, R. M. *J. Phys. Chem.* **2001**, *105*, 10388−10397.

(5) Marcus, R. A.; Sutin, N. *Biochim. Biophys. Acta* **1985**, *811*, 265−322.

(6) Levy, R. M.; Belhadj, M.; Kitchen, D. B. *J. Chem. Phys.* **1991**, *95*, 3627−3633.

(7) Figueirido, F.; Del Buono, G.; Levy, R. M. *Biophys. Chem.* **1994**, *51*, 235−241.

(8) Lee, F. S.; Chu, Z. T.; Bolger, M. B.; Warshel, A. *Protein Eng.* **1992**, *5*, 215−228.

(9) Hansson, T.; Åqvist, J. *Protein Eng.* **1995**, *8*, 1137−1144.

(10) Sham, Y. Y.; Chu, Z. T.; Tao, H.; Warshel, A. *Proteins: Struct., Funct., Genet.* **2000**, *39*, 393−407.

(11) Chen, X.; Tropsha, A. *J. Chem. Theory Comput.* **2006**, *2*, 1435−1443.

(12) Carlsson, J.; Andér, M.; Nervall, M.; Åqvist, J. *J. Phys. Chem. B* **2006**, *110*, 12034−12041.

(13) De Clercq, E. *Antiviral Res.* **1998**, *38*, 153−179.

(14) Coffin, J. M.; Hughes, S. H.; Varmus, H. E. *Retroviruses at the National Center for Biotechnology Information;* Cold Spring Harbor: Cold Spring Harbor Laboratory: 1997.

(15) Janssen, P. A. J.; Lewi, P. J.; Arnold, E.; Daeyaert, F.; de Jonge, M.; Heeres, J.; Koymans, L.; Vinkers, M.; Guillemont, J.; Pasquier, E.; Kukla, M.; Ludovici, D.; Andries, K.; de Bèthune, M.-P.; Pawels, R.; Das, K.; Clark Jr., A. D.; Frenkel, Y. V.; Hughes, S. H.; Medaer, B.; De Knaep, F.; Bohets, H.; De Clerck, F.; Lampo, A.; Williams, P.; Stoffels, P. *J. Med. Chem.* **2005**, *48*, 1901−1909.

(16) Heeres, J.; de Jonge, M.; Koymans, L. M. H.; Daeyaert, F. F. D.; Vinkers, M.; Van Aken, K. J. A.; Arnold, E.; Das, K.; Kilonda, A.; Hoornaert, G. J.; Compernolle, F.; Cegla, M.; Azzam, R. A.; Andries, K.; de Bèthune, M.-P.; Azjin, H.; Pauwels, R.; Lewi, P. J.; Janssen, P. A. J. *J. Med. Chem.* **2005**, *48*, 1910−1918.

(17) Das, K.; Clark, A. D., Jr.; Lewi, P. J.; Heeres, J.; De Jonge, M. R.; Koymans, L. M.; Vinkers, H. F.; Daeyaert, F.; Ludovici, D. W.; Kukla, M. J.; De Corte, B.; Kavash, R. W.; Ho, C. Y.; Ye, H.; Lichtenstein, M. A.; Andries, K.; Pauwels, R.; De Bethune, M. P.; Boyer, P. L.; Clark, P.; Hughes, S. H.; Janssen, P. A.; Arnold, E. *J. Med. Chem.* **2004**, *47*, 2550−2560.

(18) Das, K.; Lewi, P. J.; Hughes, S. H.; Arnold, E. *Prog. Biophys. Mol. Biol.* **2005**, *88*, 209−231.

(19) Rizzo, R. C.; Udier-Blagović, M.; Wang, D.-P.; Watkins, E. K.; Kroeger Smith, M. B.; Smith, R. H., Jr.; Tirado-Rives, J.; Jorgensen, W. L. *J. Med. Chem.* **2002**, *45*, 2970−2987.

(20) Kroeger Smith, M. B.; Hose, B. M.; Hawkins, A.; Lipchock, J.; Farnsworth, D. W.; Rizzo, R. C.; Tirado-Rives, J.; Arnold, E.; Zhang, W.; Hughes, S. H.; Jorgensen, W. L.; Michejda, C. J.; Smith, R. H., Jr. *J. Med. Chem.* **2003**, *46*, 1940−1947.

(21) Himmel, D. M.; Sarafianos, S. G.; Dharmasena, S.; Hossain, M. M.; McCoy-Simandle, K.; Ilina, T.; Clark Jr., A. D.; Knight, J. L.; Julias, J. G.; Clark, P. K.; Krogh-Jespersen, K.; Levy, R. M.; Hughes, S. H.; Parniak, M. A.; Arnold, E. **2006**, submitted for publication.

(22) Florián, J.; Goodman, M. F.; Warshel, A. *J. Phys. Chem. B* **2002**, *106*, 5739−5753.

(23) Gallicchio, E.; Kubo, M. M.; Levy, R. M. *J. Phys. Chem. B* **2000**, *104*, 6271−6285.

(24) Pierotti, R. A. *Chem. Rev.* **1976**, *76*, 717−26.

(25) Lum, K.; Chandler, D.; Weeks, J. D. *J. Phys. Chem. B* **1999**, *103*, 4570−4577.

(26) Hummer, G.; Garde, S.; García, A. E.; Paulaitis, M. E.; Pratt, L. R. *Phys. Chem. B* **1998**, *102*, 10469−82.

(27) Carlson, H. A.; Jorgensen, W. L. *J. Phys. Chem.* **1995**, *99*, 10667−10673.

(28) Warshel, A.; Russell, S. T. *Q. Rev. Biophys.* **1984**, *17*, 283−422.

(29) Åqvist, J.; Hansson, T. *J. Phys. Chem.* **1996**, *100*, 9512−9521.

(30) Cortis, C. M.; Friesner, R. A. *J. Comput. Chem.* **1997**, *18*, 1591−1608.

(31) Rocchia, W.; Sridharan, S.; Nicholls, A.; Alexov, E.; Chiabrera, A.; Honig, B. *J. Comput. Chem.* **2002**, *23*, 128−137.

(32) Lazaridis, T.; Karplus, M. *Curr. Opin. Struct. Biol.* **2000**, *10*, 139−145.

(33) Bashford, D.; Case, D. A. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129−152.

(34) Gallicchio, E.; Zhang, L. Y.; Levy, R. M. *J. Comput. Chem.* **2002**, *23*, 517−529.

(35) Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2004**, *25*, 479−499.

(36) Pratt, L. R.; Chandler, D. *J. Chem. Phys.* **1977**, *67*, 3683−3704.

(37) Huang, D. M.; Chandler, D. *J. Phys. Chem. B* **2002**, *106*, 2047−2053.

(38) Ashbaugh, H. S.; Kaler, E. W.; Paulaitis, M. E. *J. Am. Chem. Soc.* **1999**, *121*, 9243−9244.

(39) *CRC Handbook of Chemistry and Physics*, 86th ed. ed.; Lide, D. E., Ed.; CRC Press: Boca Raton, FL, 2005.

(40) Tounge, B. A.; Reynolds, C. H. *J. Med. Chem.* **2003**, *46*, 2074−2082.

(41) Singh, P.; Mhaka, A. M.; Christensen, S. B.; Gray, J. J.; Denmeade, S. R.; Isaacs, J. T. *J. Med. Chem.* **2005**, *48*, 3005−3014.

(42) Almlöf, M.; Åqvist, J.; Smålas, A. O.; Brandsdal, B. O. *Biophys. J.* **2006**, *90*, 433−442.

(43) Zhang, L.; Gallicchio, E.; Friesner, R. A.; Levy, R. M. *J. Comput. Chem.* **2001**, *22*, 591−607.

(44) Almlöf, M.; Brandsdal, B. O.; Åqvist, J. *J. Comput. Chem.* **2004**, *25*, 1242−1254.

(45) Still, W. C. Tempczyk, A.; Hawley, R. C.; Hendrikson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.

(46) Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578−1599.

(47) Onufriev, A.; Bashford, D.; Case, D. A. *J. Phys. Chem. B* **2000**, *104*, 3712−3720.

(48) Schaefer, M.; Froemmel, C. *J. Mol. Biol.* **1990**, *216*, 1045−1066.

(49) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 19824−19839.

(50) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.

(51) Jorgensen, W. L.; Madura, J. D. *Mol. Phys.* **1985**, *56*, 1381.

(52) Simonson, T. *Curr. Opin. Struct. Biol.* **2001**, *11*, 243−252.

(53) Rizzo, R. C.; Tirado-Rives, J.; Jorgensen, W. L. *J. Med. Chem.* **2001**, *44*, 145−154.

(54) Ho, W.; Kukla, M. J.; Breslin, H. J.; Ludovici, D. W.; Grous, P. P.; Diamond, C. J.; Miranda, M.; Rodgers, J. D.; Ho, C. Y.; De Clercq, E.; Pauwels, R.; Andries, K.; Janssen, M. A. C.; Janssen, P. A. J. *J. Med. Chem.* **1995**, *38*, 794−802.

(55) Smith Jr., R. H.; Jorgensen, W. L.; Tirado-Rives, J.; Lamb, M. L.; Janssen, P. A. J.; Michejda, C. J.; Kroeger Smith, M. B. *J. Med. Chem.* **1998**, *41*, 5272−5286.

(56) Hopkins, A. L.; Ren, J.; Esnouf, R. M.; Willcox, B. E.; Jones, E. Y.; Ross, C.; Miyasaka, T.; Walker, R. T.; Tanaka, H.; Stammers, D. K.; Stuart, D. I. *J. Med. Chem.* **1996**, *39*, 1589−1600.

(57) Das, K.; Ding, J.; Hsiou, Y.; Clark A. D., Jr.; Moereels, H.; Koymans, L.; Andries, K.; Pauwels, R.; Janssen, P. A.; Boyer, P. L.; Clark, P.; Smith, R. H., Jr.; Kroeger Smith, M. B.; Michejda, C. J.; Hughes, S. H.; Arnold, E. *J. Mol. Biol.* **1996**, *264*, 1085−1100.

(58) Schrödinger, Inc., Portland, OR.

(59) Banks, J. L.; Beard, J. S.; Cao, Y.; Cho, A. E.; Damm, W.; Farid, R.; Felts, A. K.; Halgren, T. A.; Mainz, D. T.; Maple, J. R.; Murphy, R.; Philipp, D. M.; Repasky, M. P.; Zhang, L. Y.; Berne, B. J.; Friesner, R. A.; Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2005**, *26*, 1752−1780.

(60) Smith, M. B. K.; Rouzer, C. A.; Taneyhill, L. A.; Smith, N. A.; Hughes, S. H.; Boyer, P. L.; Janssen, P. A. J. Moereels, H.; Koymans, L.; Arnold, E.; Ding, J.; Das, K.; Zhang, W.; Michejda, C. J.; Smith, R. H., Jr. *Protein Sci.* **1995**, *4*, 2203−2222.

(61) Ghosh, A.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B* **1998**, *102*, 10983−10990.

(62) Cheng, Y.-C.; Prusoff, W. H. *Biochem. Pharmacol.* **1973**, *22*, 3099−3108.

(63) Paulsen, M. D.; Ornstein, R. L. *Protein Eng.* **1996**, *9*, 567−571.

(64) Wang, W.; Wang, J.; Kollman, P. A. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 395−402.

(65) Wang, J.; Dixon, R.; Kollman, P. A. *Proteins Struct. Funct. Genet.* **1999**, *34*, 69−81.

(66) Su, Y.; Gallicchio, E.; Levy, R. M. *Biophys. Chem.* **2004**, *109*, 251−260.

(67) Levy, R. M.; Zhang, L. Y.; Gallicchio, E.; Felts, A. K. *J. Am. Chem. Soc.* **2003**, *25*, 9523−9530.

(68) Weeks, J. D.; Chandler, D.; Andersen, H. C. *J. Chem. Phys.* **1971**, *54*, 5237−47.

CT600258E