New Linear Interaction Method for Binding Affinity Calculations Using a Continuum Solvent Model

Ruhong Zhou,*,† Richard A. Friesner,‡ Avijit Ghosh,†,§ Robert C. Rizzo,|| William L. Jorgensen,|| and R. M. Levy⊥

IBM Thomas J. Watson Research Center, Route 134 and PO Box 218, Yorktown Heights, New York 10598, Department of Chemistry and Center for Biomolecular Simulation, Columbia University, New York, New York 10027, Department of Chemistry, Yale University, New Haven, Connecticut 06520, and Department of Chemistry, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854

Received: April 18, 2001; In Final Form: June 28, 2001

A new linear interaction energy (LIE) method based on a continuum solvent surface generalized Born (SGB) model is proposed for protein−ligand binding affinity calculations. The new method SGB-LIE is about 1 order of magnitude faster than previously published LIE methods based on explicit solvents. It has been applied to several binding sets: HEPT analogues binding to HIV-1 reverse transcriptase (20 ligands), sulfonamide inhibitors binding to human thrombin (seven ligands), and various ligands binding to coagulation factor Xa (eight ligands). The SGB-LIE predictions and cross-validation results show that about 1.0 kcal/mol accuracy is achievable for binding sets with as many as 20 ligands, e.g., for the HIV-1RT binding set, RMS errors of 1.07 and 1.20 kcal/mol are achieved for LIE fitting and leave-one-out cross validation, respectively, with correlation coefficients r^2 equal to 0.774 and 0.717. We have also explored various techniques for the LIE underlying conformation space sampling, including molecular dynamics and hybrid Monte Carlo methods, and the final results show that comparable binding energies can be obtained no matter which sampling technique is used.

1. Introduction

A central challenge in structure based rational drug design is the estimation of binding affinities for ligand−receptor complexes. A great deal of effort has been invested in this area from both academic groups and the pharmaceutical industry,[1−7] and several approaches have been developed, ranging from rapid QSAR based scoring function[1−4] to computationally intensive free energy perturbation (FEP) calculations,[5−7] but a fully satisfactory solution has not yet been developed. The rapid QSAR type approaches typically contain many approximations and usually have large errors in binding energy predictions. The more rigorous FEP approaches are more accurate but typically require little variations in the ligand structures and also demand substantial CPU costs.

The linear interaction energy approximation (LIE) is a way of combining molecular mechanics calculations with experimental data to build a model scoring function for the evaluation of ligand−protein binding free energies. LIE type methods were first suggested by Aqvist,[8−11] based upon approximating the charging integral in the free energy perturbation formula with a mean value approach in which the integral is represented as half the sum of the values at the end points, namely, the free and bound states of the ligand. Since then, the LIE method has been pursued by a number of research groups with promising results for a number of ligand binding data sets.[12−20] There is an earlier similar approach, named Linear Response Approximation (LRA), proposed by Warshel and co-workers.[21] The difference between LIE and LRA is that the LIE method neglects the electrostatic contribution over trajectories from the ligand's nonpolar state (or uncharged state, all ligand atoms' partial charge are set to zero).[22−25] The method we are going to propose in this paper is based on Aqvist's LIE approach.

From a computational standpoint, the LIE method has a number of highly attractive features. In contrast to free energy perturbations, where a large number of intermediate windows must be evaluated, the LIE method only requires simulations of the two ending windows, the ligand in pure solvent (free state) and the ligand bound to the solvated receptor (bound state). The idea is that one views the binding event as a replacement of the aqueous environment of the ligand with a mixed aqueous/protein environment. In contrast to FEP, one can study disparate ligands as long as they have similar binding modes. With FEP, only small changes between ligands are computationally tractable; the differences even in the data sets we have examined up to this point are much more significant.

Only interactions between the ligand and either the protein or the aqueous environment enter into the quantities that are accumulated during the simulation; the protein−protein and protein−water interactions are part of the "reference" Hamiltonian and hence are used to generate conformations in the simulation, via either Monte Carlo (MC) or molecular dynamics (MD), but are not used as descriptors in the resulting model for the binding free energy. This eliminates a considerable amount of noise and systematic uncertainties in the calculations, for example, arising from different conformations of the protein obtained from cocrystallized structures of different ligands.

* To whom correspondence should be addressed. E-mail: ruhongz@us.ibm.com.
† IBM Thomas J. Watson Research Center.
‡ Department of Chemistry and Center for Biomolecular Simulation.
§ Current address: Department of Computer Science, Cornell University, Ithaca, NY 14853.
|| Department of Chemistry, Yale University.
⊥ Department of Chemistry, Rutgers.

10.1021/jp011480z CCC: $20.00
Published on Web 09/21/2001

The method as implemented by Jorgensen et al.[13] contains three terms in the empirical formula, electrostatic, van der Waals, and solvent accessible surface area (SASA), for the binding energy:

$$\Delta G = \alpha(\langle U_{vdw}{}^{b}\rangle - \langle U_{vdw}{}^{f}\rangle) + \beta(\langle U_{elec}{}^{b}\rangle - \langle U_{elec}{}^{f}\rangle) + \gamma(\langle U_{SASA}{}^{b}\rangle - \langle U_{SASA}{}^{f}\rangle) \quad (1)$$

where $\langle ... \rangle$ means ensemble averages from Monte Carlo or molecular dynamics simulations. All terms are evaluated only for interactions between the ligand and its "environment", with "*f*" representing the free state and "b" representing the bound state, and α, β, and γ are LIE fitting parameters. Aqvist et al.[8−10] used only two terms in their original work, i.e., electrostatic and van der Waals interaction; however, Jorgensen et al.[13] found that it is necessary to add in one more terms for larger data sets. In our implementation discussed later, the third term is naturally replaced by the cavity energy in the continuum solvent model.

If the linear response approximation was rigorously valid, the coefficient of the electrostatic term β would be 0.5, corresponding to the mean value approximation to the charging integral. In fact, one can recover a value very close to this for less complex systems, for example solvation of small molecules in water. However, some of the steps involved in the binding event, such as the removal of water from the protein cavity and subsequent introduction of the ligand, are unlikely to be accurately described by a fully linear model. Therefore, in practice, optimization of fitting parameters yields the electrostatic coefficients that are significantly different from the ideal value of 0.5.[13] By allowing this empirical element, one is sacrificing generality: the method probably requires the ligands to have similar binding modes, and new parameters must be developed for each receptor. In return, however, one can obtain a reasonable level of accuracy (reflected in cross-validation studies as well as the overall fitting accuracy, see below) with a relatively modest expenditure of CPU time, under assumptions that are quite reasonable for many structure based drug design projects. Our new approach based on the continuum solvent model, surface generalized Born (SGB) model, is more than 1 order of magnitude faster than the methods based on the explicit solvent models with comparable accuracy in binding energy prediction.

2. Methodology

We have developed an implementation of the LIE, in the context of the IMPACT program of Levy and co-workers,[26,27] using the surface generalized Born (SGB) continuum solvation model[28] and the OPLS-AA force field of Jorgensen and co-workers.[29] To our knowledge, this is the first implementation of LIE based on continuum solvation models, for simplicity we called it SGB-LIE in the following context. Key features of the implementation are as follows.

First, we replaced the solvent accessible surface area term in Jorgensen's LIE formulation by the cavity term in continuum solvent model:

$$\Delta G = \alpha(\langle U_{vdw}{}^{b}\rangle - \langle U_{vdw}{}^{f}\rangle) + \beta(\langle U_{elec}{}^{b}\rangle - \langle U_{elec}{}^{f}\rangle) + \gamma(\langle U_{cav}{}^{b}\rangle - \langle U_{cav}{}^{f}\rangle). \quad (2)$$

The use of a continuum model provides much more rapid convergence of the simulations and at the same time higher accuracy in terms of treatment of the long-range electrostatic forces: for example, there is no need to keep the protein system neutral as is done in explicit solvent models to avoid the Born correction due to the finite size of the solvent sphere[8,13] (typically 20 Å from the active site). The statistics on the various interaction terms (total four possible terms now, van der Waals energy and Coulomb energy between the ligand and the protein, and reaction field energy and cavity energy between the ligand and the solvent, see below for details) are better converged than in an explicit solvent simulation, and the required CPU time is much smaller.

In the generalized Born models, there are two so-called reaction field energy and cavity energy terms in the total sum of the solvation free energy,[30,28]

$$U_{SGB} = U_{rxn} + U_{cav} \quad (3)$$

and there is no explicit electrostatic energy and van der Waals energy between solute and solvent any more. The van der Waals energy is implicitly included in the cavity term, while the reaction field energy (or charging free energy) is half the Coulombic energy between solute and solvent if the linear interaction approximation is exact, which is proved to be the case for small molecules' solvation in Jorgensen's paper;[12] thus, we use

$$U_{elec} = U_{coul} + 2U_{rxn} \quad (4)$$

as our total electrostatic energy term, where U_{coul} stands for the possible Coulomb interaction between ligand and protein. This gives total four possible LIE interaction components, van der Waals energy and Coulomb energy (between ligand and protein), and reaction field energy and cavity energy (between ligand and continuum solvent). The Coulomb energy and reaction field energy are combined as a total electrostatic energy in SGB-LIE, as shown in eq 4.

The calculation of the reaction field energy for the ligand in the free state is straightforward.[28] However, the reaction field energy between ligand and continuum solvent in the bound state is more complicated to calculate, since SGB generally reports the total reaction field energy for the whole solute, i.e., for the ligand and protein together, although the reaction field energy from the protein is not needed in our SGB-LIE energy term. The equation for the total reaction field energy in SGB is expressed as[28]

$$U_{rxn} = \sum_i U_{se}(q_i,\mathbf{r}_i) + \sum_{i\neq j} U_{pr}(q_i, q_j,\mathbf{r}_i,\mathbf{r}_j), \quad (5)$$

where the single energy U_{se} is

$$U_{se} = -\frac{1}{8\pi}(1 - 1/\epsilon)\int_S \frac{q_k{}^2}{|\mathbf{R} - \mathbf{r}_k|^4}(\mathbf{R} - \mathbf{r}_k)\cdot\mathbf{n}(\mathbf{r})\, d^2\mathbf{R}, \quad (6)$$

and the pairwise screened Coulomb energy is

$$U_{pr} = -\frac{1}{2}(1 - 1/\epsilon)\frac{q_i q_j}{\sqrt{r_{ij}^2 + \alpha_{ij}^2 e^{-D}}}, \quad (7)$$

with parameter $\alpha_{ij} = \sqrt{\alpha_i \alpha_j}$ (α_i and α_j are the Born radius) and parameter $D = r_{ij}^2/(2\alpha_{ij})^2$. We therefore count the single energies U_{se} from ligand atoms only, and the pairwise screened Coulombic energies U_{pr} as a whole if both atoms are from the ligand, and half if one from the ligand and the other from the protein, and zero if both from the protein.

The efficiency of the original SGB model by Ghosh et al.[28] has been dramatically improved by utilizing multiple time step techniques[31,32] in SGB. After carefully analyzing the different time scales in SGB's force terms, mainly the single energy term and the pairwise screened Coulomb term (the cavity term is very fast), we found that the single energy term has a much longer time scale compared to the pairwise Coulombic interactions. Therefore, it is ideal to use the multiple time step algorithms to update Coulomb interactions much more often than the single energy terms. In addition, the pairwise screened Coulomb interactions are further decomposed into short- and long-range contributions based on pair distance, with different updating frequencies. Detailed implementations and speedups will be described elsewhere.[33]

Two sampling techniques, molecular dynamics (MD) and hybrid Monte Carlo (HMC), have been used for LIE conformation space sampling in the present work. All simulations are carried out with the IMPACT package. A conjugate gradient minimization is performed first, starting from the initial docked structures, and then a 15 ps MD equilibration is followed with temperature smoothly increasing from 0 to 310 K by velocity scaling and resampling, finally a 15−30 ps MD simulation is run for the SGB-LIE data collection. The MD time step used is 2.0 fs unless otherwise explicitly stated, with the use of the multiple time step algorithm RESPA.[31,32] No cutoff is used for the long-range electrostatic interactions in the simulations, and the SGB solvation is called every 10 steps of MD, with a dielectric constant of 1.0 for the solute and 80.0 for the water solvent. The clipped residues on the boundary are capped with acetyl (ACE) and methylamine (NMA) groups (some proteins, like HIV-1RT, are too big for this type of simulations, so residues greater than 20 Å away from the active site are typically truncated).[10,12,14] During MD simulations, these capped residues are held in a buffer region using a harmonic constraint potential with a force constant of 25 kcal/mol.

3. Results and Discussion

Three receptor−ligand binding sets are used in this work for testing the new SGB-LIE method: HEPT analogues binding to HIV-1RT, sulfonamide analogues binding to human thrombin, and various ligands binding to coagulation factor Xa. The initial system setups for the HIV-1RT binding set and thrombin binding set are the same as previously reported,[13,14] and the initial docked structures for the coagulation factor Xa set is obtained from an industrial partner (see below). Results for these binding sets, including cross-validation studies, are presented, and in each case, the predictive capability of the model is quite reasonable, while the CPU time is within an acceptable range for giving experimental chemists rapid feedback.

With three fitting parameters in the LIE model, it is generally true that the smaller the binding set, the easier the fitting and the better the results. Most of the early LIE works have less than 10 ligands in their binding sets.[8−10,13,22] In the three binding sets studied here, the first one (HIV-1RT) has 20 ligands, the second one (thrombin) has seven ligands, and the third one (factor Xa) has eight ligands. We select the first one, which has the most ligands, to be our show case as to demonstrate the continuum solvent based SGB-LIE method. Also, this is a very interesting and important set of ligands for HIV/AIDS treatments. Currently, the use of nucleoside inhibitors, nonnucleoside inhibitors, and HIV protease inhibitors (and/or combinations) is the best method for controlling the HIV infection, and this binding set, HEPT analogues, belongs to the nonnucleoside inhibitor class. In fact, one of the ligands, MKC-442 (H11, see Table 1) is in clinical trials.[34,35]

TABLE 1: Molecular Structures of HEPT Analogs: Details of R1, R2, and R3 Groups in Figure 1[a]

number	R1	R2	R3	EC_{50}	ΔG_{bind}
H01	Me	$CH_2OCH_2CH_2OH$	SPh	7.0	−7.32
H02	Me	$CH_2OCH_2CH_2CH_3$	SPh	3.6	−7.73
H03	Me	$CH_2OCH_2CH_3$	SPh	0.33	−9.20
H04	Me	CH_2OCH_3	SPh	2.1	−8.06
H05	Me	CH_2OCH_2Ph	SPh	0.088	−10.01
H06	*i*-Pr	CH_2OCH_2Ph	SPh	0.0027	−12.16
H07	Me	Et	SPh	2.2	−8.03
H08	Me	Me	SPh	>150	> −5.43
H09	Et	$CH_2OCH_2CH_3$	SPh	0.019	−10.96
H10	*i*-Pr	$CH_2OCH_2CH_3$	SPh	0.012	−11.24
H11	*i*-Pr	$CH_2OCH_2CH_3$	CH_2Ph	0.004	−11.89
H12	*c*-Pr	$CH_2OCH_2CH_3$	SPh	0.1	−9.93
H13	Me	$CH_2OCH_2CH_2OH$	CH_2Ph	23.0	−6.52
H14	Me	$CH_2OCH_2CH_2OH$	OPh	85.0	−5.78
H15	Me	$CH_2OCH_2CH_2OH$	SPh-3,5 di-Me	0.26	−9.35
H16	Et	$CH_2OCH_2CH_2OH$	SPh-3,5 di-Me	0.013	−11.19
H17	*i*-Pr	$CH_2OCH_2CH_2OH$	SPh-3,5 di-Me	0.0027	−12.16
H18	Et	CH_2OCH_2Ph	SPh	0.0059	−11.68
H19	Me	H	SPh	>250	> −5.11
H20	Me	Bu	SPh	1.2	−8.40

[a] The experimental activities are taken from works done by Tanaka et al.,[36−39] EC_{50} in μM at 310 K, and $\Delta G_{bind} \approx RT \ln(EC_{50})$ in kcal/mol.

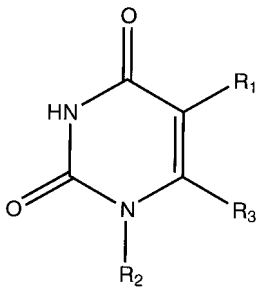

Figure 1. Molecular structures of HEPT analogues binding to HIV-1 reverse transcriptase. See Table 1 for details of R1, R2, and R3 groups.

The molecular structure of the HEPT analogues of the first binding set is shown in Figure 1, with the three substitutes on the uracil ring, R1, R2 and R3 groups listed in Table 1. In the simplified notations used in Table 1, "Me" stands for the methyl group, and "Et" for the ethyl group, "Pr" for the propyl group, "Bu" for the butyl group, and "Ph" for the phenyl group. As we can see from the table, the differences in these analogues are quite significant from the stand point of FEP calculations. In a typical FEP simulation, the differences in various ligands under perturbation are normally very small to avoid too many windows.

The experimental activities[36−39] are also listed in Table 1. With only three fitting parameters, it is a nontrivial task to predict the binding affinities for all the 20 ligands. As mentioned earlier, many of the previous LIE works have much less ligands in their binding sets.[8−10,13,20,22] The protein−ligand system setup is identical to those used in a previous study,[16] i.e., starting from the initial PDB structure 1rt1 for docked MKC-442 (H11), all the other ligands' binding structures are built based on this template. The protein HIV-1RT size has been reduced by truncating distant residues from the active site to save the computational cost. The final protein model has 123 residues and 1923 atoms. See Figure 2 for a sketch of the protein system and the docked ligand H01, and consult the previous paper[16] for more details in the initial setup. A conjugate gradient

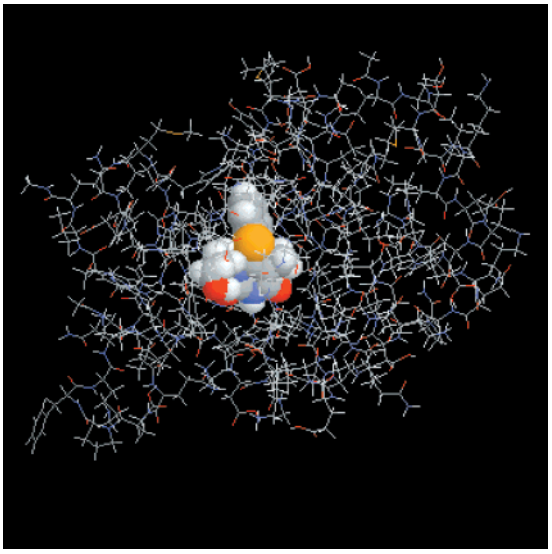

Figure 2. Structure of the docked HEPT (H01) inside the active site of HIV-1RT. The ligand is represented with space-fill option.

minimization is then performed with SGB continuum solvation turned on, followed by a 15 ps MD equilibration. Finally a 15−30 ps MD data collection is run. No cutoff is used in the long-range electrostatic interactions. A dielectric constant of 1.0 is used for solute (protein and ligand) and 80.0 for water in continuum solvent SGB model. This procedure was repeated for all the binding sets studied here.

As explained in section 2, there are in total four possible LIE interaction components in our SGB-LIE method: Coulomb energy, van der Waals energy, reaction field energy, and cavity energy. The Coulomb and van der Waals interactions between ligand and environment are zero in the free state, since the solvent is represented by the continuum solvation completely; thus, the only interactions available for the free state are the reaction field energy and cavity energy, while in the bound state, there are Coulomb and van der Waals interactions between ligand and protein, and there are also reaction field energy and cavity energy between ligand and continuum solvent. To determine how long the MD simulation is needed for the four LIE components to converge, we have plotted the time averages for each component vs MD time for HEPT analogue H01 in Figure 3. The cavity energy and van der Waals energy converge faster than the Coulomb energy and reaction field energy, which makes sense since the electrostatic interactions are the long-range interactions and converge much slowly, while the van der Waals interactions are typically regarded as short-range forces and die off quickly, and the cavity energy is normally based on the solvent accessible surface area which depends on the overall molecular shape and typically converges very fast. In this particular case, the cavity energy converges for H01 in less than 5 ps in both the free and bound states, and the van der Waals energy converges in about 10 ps, the reaction field energy converges in less than 15 ps in both the free and bound states, and the long-range Coulomb energy also converges fairly well after 15 ps. We have also run other HEPT analogues for longer time, and the conclusion is roughly the same; that is, for this ligand−receptor binding set, a 15 ps MD data collection run yields reasonably converged results. For other binding sets, we have used 20−30 ps MD simulations (see below).

Table 2 lists the time averages of the SGB-LIE interaction components for the HIV-1RT binding set from 15 ps MD data collection. Using the three-parameter model detailed in section 2, a least-squares fit based on singular value decomposition

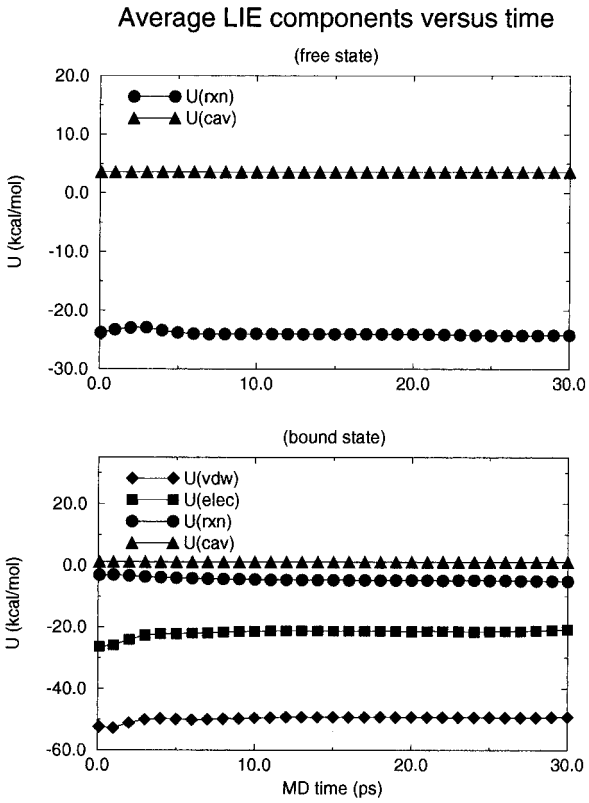

Figure 3. The variation of the average LIE interaction components vs MD time. The results are obtained from a 30 ps MD simulation following a 15 ps MD equilibration for HEPT binding with HIV-1RT. The top one shows the averages for the free state, and the bottom one for the bound state.

(SVD) is performed to the 20 ligand−receptor binding affinities. The final results are shown in Figure 4, which plots the SGB-LIE binding energy predictions vs the experimental values. If a predicted binding energy agrees exactly with the experimental value, a data point (represented by diamonds) exactly on the diagonal line will be shown. To help visualizing these data points, a lower and upper bound line are also plotted in the figure, with 1.0 kcal/mol below or above the experimental values. From the figure, most of the data points (19 out of 20) are within or very close to these two bound lines, which means most of them have either less than or about 1.0 kcal/mol error. The only data point that shows large deviations from the experimental value is ligand H11 (MKC-442), which has 2.58 kcal/mol error. The reason for this large error for H11 case is not immediately clear, although it might be possible that the experimental potency of H11 is overestimated (in all the simulations done here and some simulations done with explicit solvent model, this H11 always shows a large deviation from the experimental value with LIE predictions less potent than the experimental affinity, see below). One evidence for this experimental overestimation might be that this MKC-442's (H11) binding energy is 2.24 kcal/mol more potent than that of nevirapine (nevirapine is another nonnucleoside inhibitor and is one of the few drugs in this class approved by FDA for HIV/AIDS treatments. It is not studied here, but was included in our earlier paper[14]), but another study showed that MKC-442 (H11) is only about 1.0 kcal/mol more potent than nevirapine.[40] Thus, if we use nevirapine as a reference point, MKC-442's potency might be overestimated by 1.2 kcal/mol in the data we adopted. Of course, the differences might be from other sources and also the error could be in the binding affinity for nevirapine;

TABLE 2: Time Averages of LIE Interaction Components for the HIV-1RT Binding Set[a]

ligand	$\langle U_{coul}^f \rangle$	$\langle U_{vdw}^f \rangle$	$\langle U_{rxn}^f \rangle$	$\langle U_{cav}^f \rangle$	$\langle U_{coul}^b \rangle$	$\langle U_{vdw}^b \rangle$	$\langle U_{rxn}^b \rangle$	$\langle U_{cav}^b \rangle$
H01	0.0	0.0	−24.317	3.625	−25.235	−48.740	−2.945	1.099
H02	0.0	0.0	−20.927	3.708	−12.886	−51.788	−5.794	1.103
H03	0.0	0.0	−20.692	3.580	−14.263	−49.583	−6.247	1.096
H04	0.0	0.0	−20.502	3.434	−12.626	−48.672	−5.380	1.105
H05	0.0	0.0	−19.600	3.959	−17.141	−58.687	−2.740	1.107
H06	0.0	0.0	−20.110	4.228	−17.568	−62.340	−4.715	1.112
H07	0.0	0.0	−17.863	3.340	−13.775	−46.880	−4.215	1.095
H08	0.0	0.0	−18.874	3.224	−12.528	−45.175	−3.534	1.096
H09	0.0	0.0	−21.572	3.715	−17.239	−52.929	−6.843	1.100
H10	0.0	0.0	−20.497	3.737	−15.964	−51.868	−6.696	1.095
H11	0.0	0.0	−20.717	3.710	−14.396	−52.074	−6.048	1.100
H12	0.0	0.0	−21.555	3.774	−14.019	−53.019	−5.709	1.104
H13	0.0	0.0	−24.239	3.576	−22.580	−47.590	−4.035	1.097
H14	0.0	0.0	−23.174	3.513	−20.229	−49.273	−5.702	1.093
H15	0.0	0.0	−24.946	3.933	−21.328	−55.545	−5.453	1.097
H16	0.0	0.0	−25.558	3.972	−23.490	−58.115	−6.601	1.098
H17	0.0	0.0	−24.012	4.061	−20.338	−60.584	−5.567	1.096
H18	0.0	0.0	−20.562	4.080	−15.895	−60.592	−5.822	1.101
H19	0.0	0.0	-21.665	3.132	-15.856	-44.921	-4.068	1.092
H20	0.0	0.0	-18.785	3.615	-16.408	-49.398	-4.973	1.102

[a] All energies are in kcal/mol. The data are collected from a 15 ps MD simulation after a 15 ps MD equilibration. Each ligand shows the LIE components for both free and bound states, with "*f*" denoting the free state and "b" denoting the bound state.

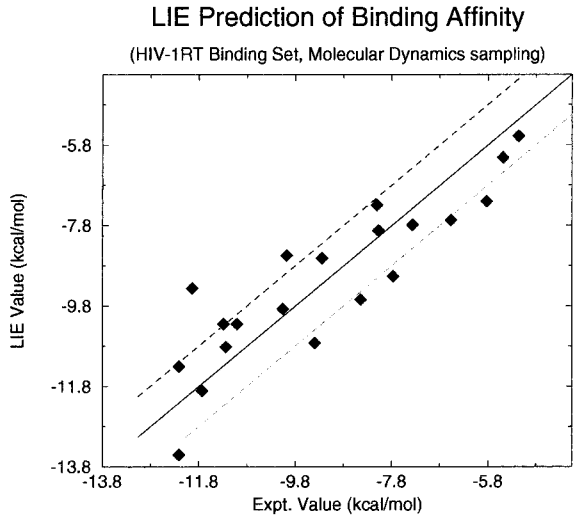

Figure 4. LIE binding energies for the HIV-1RT binding set from MD sampling. The overall RMS error is 1.07 kcal/mol for 20 ligands studied here. If LIE results agree perfectly with the experimental values, the data points (represented by diamonds) should be on the diagonal line.

anyway, we think more experimental studies might be worthwhile here. Also, uncertainties are not reported in these experiments, but they are typically at least 0.5 kcal/mol. Interestingly, explicit solvent based LIE models found very similar results, i.e., the experimental value of H11 is larger (more negative) than that from LIE predictions (more discussions below).

The detailed numerical results are summarized in Table 3. The overall RMS deviation of the SGB-LIE predictions for this binding set is 1.07 kcal/mol, which is quite encouraging for such a simple model. The correlation coefficient r^2 is 0.776, which indicates a good correlation with experiment. The three SGB-LIE parameters are found to be $\alpha = 0.178$, $\beta = 0.320$, and $\gamma = 1.74$. Since the cavity energy formula used in SGB model is based on the total solvent accessible surface area (SASA),

$$U_{cav} = c_1 \mathrm{SASA} + c_2 \tag{8}$$

where c_1 and c_2 are empirical coefficients; with $c_1 = 0.00486$ kcal/mol Å^2 and $c_2 = 1.092$ kcal/mol,[28,41] we can easily convert

TABLE 3: LIE Fitting and Cross-Validation (Jackknife) Results for the HIV-1RT Binding Set from MD Sampling[a]

ligand	expt	LIE	jackknife
H01	−7.32	−7.73	−7.82
H02	−7.73	−9.01	−9.11
H03	−9.20	−8.56	−8.53
H04	−8.06	−7.24	−7.10
H05	−10.01	−9.82	−9.80
H06	−12.16	−13.45	−14.73
H07	−8.03	−7.87	−7.86
H08	−5.43	−6.06	−6.20
H09	−10.96	−10.20	−10.12
H10	−11.24	−10.20	−10.08
H11	−11.89	−9.31	−9.06
H12	−9.93	−8.49	−8.31
H13	−6.52	−7.62	−7.77
H14	−5.78	−7.15	−7.38
H15	−9.35	−10.67	−10.91
H16	−11.19	−10.77	−10.70
H17	−12.16	−11.25	−11.17
H18	−11.68	−11.85	−11.87
H19	−5.11	−5.53	−5.71
H20	−8.40	−9.59	−9.78
RMS error		1.07	1.31

[a] All energies are in kcal/mol. An overall RMS error of 1.07 kcal/mol is achieved for this LIE fitting, and an RMS error of 1.31 kcal/mol for the cross-validation tests.

γ to be based on the solvent accessible surface area, which can then be compared with the explicit solvent model. We call this new parameter γ_0,

$$\gamma_0 = c_1 \gamma \tag{9}$$

With this conversion, the new parameters are $\alpha = 0.178$, $\beta = 0.320$, and $\gamma_0 = 0.00847$. In all the following fittings, we report this γ_0 based on the SASA. It needs to be pointed out though that other implementations for the cavity term in continuum solvent models are more sophisticated and the coefficient c_1 can depend on atom types, so the first term in the cavity energy becomes a sum over all atoms. Therefore, this γ_0 parameter based on total SASA cannot be defined. In these cases, we can always use the original γ that is based on the total cavity energy.

Except for the analogue H11 mentioned above, the LIE calculations agree with experiments quite well. A closer look

at the components in the LIE binding energy for each ligand reveals some important points. For example, the experiments show that when the R1 group goes from Me group to Et group (H03 to H09, H15 to H16, H05 to H18), and from Et group to *i*-Pr group (H09 to H10, H16 to H17) while R2 and R3 groups stay the same, the ligand potency increases: this is also confirmed in our SGB-LIE predictions. The reason behind this is coming from the fact that the van der Waals interactions between ligand and protein increase when R1 group changes from smaller Me group to larger Et and *i*-Pr groups. Also the net cavity energy loss due to the burial of solvent accessible surface area increases from Me to Et to *i*-Pr groups. The ligand H01 and H14 only differs in group R3, from SPh to OPh, and experimental results show that there is a 1.5 kcal/mol difference, with H01 (SPh) more potent than H14 (OPh). Our SGB-LIE results agree qualitatively with the experiments, but with a smaller difference, 0.6 kcal/mol. Overall, we found the binding affinities for this binding set are largely coming from the van der Waals interactions between ligands and HIV-1RT receptor (i.e., needs a good geometric fit), and the net loss of the cavity energy which is the same as the burial of solvent accessible surface area. These findings agree well with the new five-term model proposed by Jorgensen et al.[14] based on explicit solvents. The authors used four descriptors and one constant in describing HEPT and nevirapine analogues binding to HIV-1RT. The four descriptors are (1) the change in the total number of H-bonds for the inhibitor in going from unbound to bound states, (2) the ligand−protein van der Waals interaction energy, (3) the change in hydrophobic SASA upon bindings, and (4) a secondary amide indicator. Some of the major conclusions from the four descriptor model are (1) the loss of the H-bonds with inhibitor is unfavorable, (2) burial of the hydrophobic surface area is favorable, and (3) a good geometrical fit without steric clashes is needed. Even though the descriptors used in this work and the previous one are quite different, the general conclusion for the origin of the binding affinities seems to agree very well.

In summary, this overall 1.07 kcal/mol RMS error and correlation coefficient r^2 of 0.776 indicate that our model agrees with experiments quite well, and our RMS error is also comparable with some of our preliminary results with explicit solvent model based on the same three-term LIE model. In the preliminary results with a similar three-term model, the average unsigned error is 0.99 kcal/mol with parameters $\alpha = 0.0470$, $\beta = 0.153$, and $\gamma = 0.0206$, and another separate fit shows an average unsigned error of 0.90 kcal/mol with parameters $\alpha = -0.09812$, $\beta = 0.118$ and $\gamma = 0.0258$. Interestingly, the H11 (MKC-442) case is also among the cases showing large errors, with an experimental value of about 2.0 kcal/mol being more potent than that of the LIE calculations. It should be noted that our continuum solvent based model treats the three individual energy terms differently from those in the explicit solvent model (especially the van der Waals energy), so the three parameters cannot be compared head to head, but nevertheless they are in the same ball park. In the new five-term model proposed by Jorgensen et al. for the extended LIE,[14] it also shows that MKC-442 (H11) has a much bigger error compared to other HEPT analogues (1.82 kcal/mol more potent in experimental value[14]).

A cross validation is further checked to see how well the model can predict unknown binding energies. It is done by a so-called Jackknife (or leave-1-out) test, that is to leave one ligand out of the training set, and then use the parameters obtained from the rest of the binding set (19 ligands in this case) to predict the one left out. Table 3 lists the Jackknife cross-validation results, with each ligand's binding energy being predicted from the fitting of the rest 19 ligands. The correlation coefficient r^2 is 0.662, indicating a reasonable prediction ability. The overall RMS error is 1.31 kcal/mol, which is still quite reasonable for true predictions of all 20 ligands. With the HMC sampling, the cross-validation results are slightly better, with RMS error of 1.20 kcal/mol and correlation coefficient r^2 of 0.717 (see below). The biggest error is again coming from ligand H11, with 2.74 kcal/mol error. Another significant error is from ligand H06, with 2.57 kcal/mol error. We will come back to this H06 case, since from the fitting parameters we noticed that H06's cross-validation fitting has significantly different parameters from all the other cases, including the fittings from the HMC sampling results (see below).

To test how sensitive the LIE method is to the underlying sampling techniques, or in other words, how good the sampling technique is in surfing the local conformation space, we also implemented LIE with the HMC sampling.[42,43] It should be noted that in LIE method, we are typically sampling the local conformations around the docked ligand structures, and for the proteins only the conformations near the active site is of most significance, so relatively short MD or MC simulations are enough for the LIE purpose.[9,13] If a global conformation space needs to be sampled, then neither MD nor MC is good enough in this regard, and other techniques such as Jump Walking,[44] Smart Walking,[43] Monte Carlo/minimization,[45] or other efficient sampling methods,[46,47] will be more appropriate.

The time averages of the LIE components from HMC sampling are summarized in Table 4. The HMC sampling started from the same equilibrated states as in the MD sampling described above. A Metropolis accept/reject criterion is checked every 5 step of HMC's underlying MD simulation. The time step used in HMC's underlying MD is 3.0 fs with RESPA algorithm,[31,32] slightly bigger than 2.0 fs used in the regular MD sampling, since we do not need to worry about the stability issue of the MD simulation in HMC (HMC is often called bad MD, but good MC). These time averages are obtained from a total of 7500 steps of the underlying MD with time step 3.0 fs, analogous to a "22.5 ps" MD simulation, but not completely equivalent, since it is essentially a Monte Carlo method. Because the total number of MD steps in HMC sampling are the same as in the regular MD sampling, the overall CPU cost of the two sampling techniques are comparable (the administrative cost in HMC is negligible compared to force evaluations). Using the same three-parameter model, the LIE predictions are shown in Figure 5. Again, as we can see from the figure, most of the data points (18 out of 20) are within or very close to these two bound lines, which means most of them have either less than or about 1.0 kcal/mol error. The data point H11 again has a large error, 2.62 kcal/mol, but data point H14 also shows some significant error, 2.76 kcal/mol. The reason for the big error in H14 case is not very clear: it could be due to the force field parameters used since this is the only case whose R3 group has an Oxygen atom bonded to the phenyl group (−OPh). It could also be due to the fitting procedure used since H14 was fine in the MD case. The detailed numbers are also summarized in Table 5. The overall RMS deviation is 1.07 kcal/mol and the correlation coefficient is 0.774, which are both comparable to those from MD sampling. The new parameters are found to be $\alpha = 0.262$, $\beta = 0.368$, and $\gamma_0 = 0.00213$, which are comparable to those from MD sampling. The cross-validation results are also listed in Table 5. The overall RMS error for the Jackknife test is 1.20 kcal/mol, slightly better than that in the MD sampling which was 1.31 kcal/mol. The correlation coefficient is 0.717 which is also slightly better than 0.662 in MD sampling. The

TABLE 4: Time Averages of LIE Interaction Components for the HIV-1RT Binding Set[a]

ligand	$\langle U_{coul}^f \rangle$	$\langle U_{vdw}^f \rangle$	$\langle U_{rxn}^f \rangle$	$\langle U_{cav}^f \rangle$	$\langle U_{coul}^b \rangle$	$\langle U_{vdw}^b \rangle$	$\langle U_{rxn}^b \rangle$	$\langle U_{cav}^b \rangle$
H01	0.0	0.0	−24.054	3.616	−25.950	−47.865	−2.998	1.106
H02	0.0	0.0	−20.227	3.688	−13.753	−51.258	−6.134	1.105
H03	0.0	0.0	−20.570	3.584	−14.806	−49.619	−5.966	1.097
H04	0.0	0.0	−20.266	3.424	−12.449	−47.831	−5.709	1.102
H05	0.0	0.0	−20.346	3.943	−18.137	−58.276	−2.709	1.119
H06	0.0	0.0	−18.801	4.049	−18.423	−64.133	−4.869	1.153
H07	0.0	0.0	−18.094	3.336	−15.107	−46.464	−3.816	1.101
H08	0.0	0.0	−18.639	3.221	−14.704	−43.517	−2.868	1.100
H09	0.0	0.0	−21.566	3.624	−18.048	−52.390	−7.036	1.103
H10	0.0	0.0	−21.396	3.750	−17.880	−53.077	−6.393	1.097
H11	0.0	0.0	−21.644	3.704	−16.877	−51.309	−6.383	1.103
H12	0.0	0.0	−21.872	3.759	−14.831	−54.084	−5.429	1.100
H13	0.0	0.0	−24.925	3.577	−22.803	−48.229	−5.247	1.097
H14	0.0	0.0	−24.450	3.557	−19.908	−47.722	−5.723	1.110
H15	0.0	0.0	−23.732	3.891	−23.729	−55.138	−5.595	1.099
H16	0.0	0.0	−25.123	3.984	−22.471	−56.710	−7.074	1.097
H17	0.0	0.0	−24.237	4.061	−21.744	−60.506	−6.027	1.095
H18	0.0	0.0	−20.554	4.052	−17.164	−61.182	−5.458	1.117
H19	0.0	0.0	−21.400	3.134	−16.764	−42.450	−4.324	1.107
H20	0.0	0.0	−18.408	3.625	−16.828	−48.105	−5.0003	1.097

[a] All energies are in kcal/mol. The data are collected from a 15000 steps HMC simulation after a 15 ps MD equilibration. Each ligand shows the LIE components for both free and bound States, with "*f*" denoting the free state and "b" denoting the bound state.

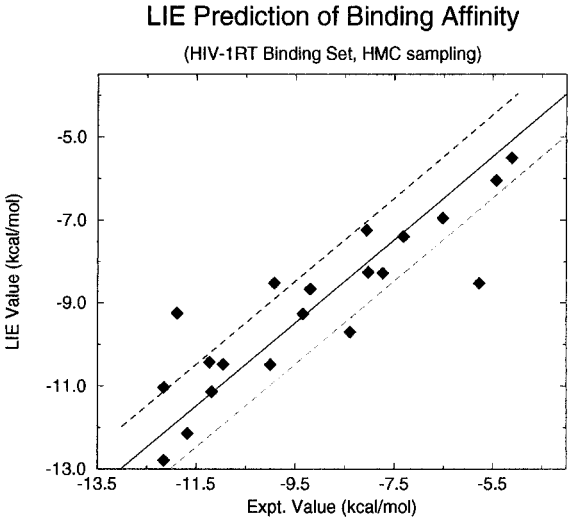

Figure 5. LIE binding energies for the HIV-1RT binding set from HMC sampling. The overall RMS error is 1.07 kcal/mol for 20 ligands studied here. If LIE results agree perfectly with the experimental values, the data points (represented by diamonds) should be on the diagonal line.

TABLE 5: LIE Fitting and Cross-Validation (Jackknife) Results for the HIV-1RT Binding Set from HMC Sampling[a]

ligand	exptl	LIE	jackknife
H01	−7.32	−7.42	−7.45
H02	−7.73	−8.30	−8.36
H03	−9.20	−8.68	−8.65
H04	−8.06	−7.27	−7.16
H05	−10.01	−10.51	−10.56
H06	−12.16	−12.81	−12.98
H07	−8.03	−8.28	−8.30
H08	−5.43	−6.07	−6.23
H09	−10.96	−10.50	−10.44
H10	−11.24	−10.45	−10.34
H11	−11.89	−9.27	−9.13
H12	−9.93	−8.54	−8.39
H13	−6.52	−6.98	−7.16
H14	−5.78	−8.54	−8.68
H15	−9.35	−9.28	−9.27
H16	−11.19	−11.16	−11.15
H17	−12.16	−11.06	−10.94
H18	−11.68	−12.16	−12.27
H19	−5.11	−5.53	−5.78
H20	−8.40	−9.72	−10.30
RMS error		1.07	1.20

[a] All energies are in kcal/mol. An overall RMS Error of 1.07 kcal/mol is achieved for this LIE fitting, and an RMS error of 1.20 kcal/mol for the cross-validation tests.

improvement over the MD sampling for the Jackknife test is from the fact that the troublesome case H06 in MD sampling is now fixed, with only 0.82 kcal/mol error in HMC sampling. Of course, the biggest error in the Jackknife test for HMC sampling is still from ligand H11, which shows a 2.76 kcal/mol error.

Since the parameters α, β, and γ are crucial in the LIE method, a natural question will arise: How close are these parameters from one fit to another fit? If the LIE model really has the ability to predict binding affinities, one might expect that the parameters should be comparable from different fittings for the same binding set. Of course, we should not expect them to be identical due to the "best possible fit" inside the SVD fitting procedure. As we have already seen from above, the parameters are indeed comparable for the LIE fitting using either MD sampling or HMC sampling. Then another question will arise: How about all the other Jackknife cross-validation fittings? There are 20 sets of them for each of the two sampling methods used. Table 6 lists all the fitting parameters from the Jackknife cross-validation tests in both MD and HMC samplings. Almost all the cases have comparable parameters, except the case H06 in MD sampling, which has a significant negative $\gamma_0 = -0.0115$ and a much larger α value ($\alpha = 0.401$). Interestingly, this is the case that shows a large error in the Jackknife test in the MD sampling, which also results a slightly higher overall RMS error as compared to that in the HMC sampling. We think the error might come from the nonstable SVD fitting, and further investigation is needed here.

We have also studied two other binding sets using our SGB-LIE model, one with seven ligands (sulfonamide analogues binding to thrombin), and the other with eight ligands (various ligands binding to coagulation factor Xa). Since they are much smaller, one might argue that they are easier to fit in LIE. It is true in general that the smaller the binding set, the easier the fit, so we only briefly show the final fitting results here. Figure 6 plots the LIE binding energies for sulfonamide and its

TABLE 6: Fitting Parameters α, β, and γ for LIE with Data from Both MD and HMC Samplings, Including All the Jackknife Cross-Validations (−H01 Means Fitting without H01 Case)[a]

	MD			HMC		
fits	α	β	γ	α	β	γ
all	0.1782	0.3199	0.0085	0.2618	0.3683	0.0021
−H01	0.1577	0.3179	0.0105	0.2548	0.3652	0.0027
−H02	0.1562	0.3225	0.0108	0.2478	0.3591	0.0033
−H03	0.1848	0.3209	0.0078	0.2659	0.3695	0.0017
−H04	0.1565	0.3401	0.0111	0.2444	0.3802	0.0041
−H05	0.1736	0.3221	0.0089	0.2778	0.3724	0.0007
−H06	0.4006	0.3804	−0.0115	0.2770	0.3901	0.0013
−H07	0.1743	0.3214	0.0089	0.2712	0.3701	0.0013
−H08	0.1951	0.2992	0.0063	0.2795	0.3522	0.0001
−H09	0.1660	0.3103	0.0093	0.2498	0.3584	0.0030
−H10	0.1969	0.3033	0.0061	0.2680	0.3545	0.0010
−H11	0.2436	0.2998	0.0012	0.2795	0.3684	0.0001
−H12	0.1580	0.3480	0.0110	0.2981	0.3908	−0.0009
−H13	0.1478	0.3084	0.0113	0.2215	0.3494	0.0056
−H14	0.1480	0.2952	0.0109	0.2922	0.3684	−0.0006
−H15	0.1329	0.3480	0.0138	0.2647	0.3694	0.0019
−H16	0.1946	0.3120	0.0066	0.2611	0.3678	0.0022
−H17	0.1750	0.3124	0.0085	0.2319	0.3569	0.0046
−H18	0.1821	0.3222	0.0082	0.2819	0.3832	0.0007
−H19	0.2008	0.2994	0.0057	0.2921	0.3535	−0.0012
−H20	0.1368	0.3436	0.0133	0.1539	0.3833	0.0133

[a] The fitting parameters are comparable across all the fittings, except for analog H06 in MD sampling (see discussions in the text).

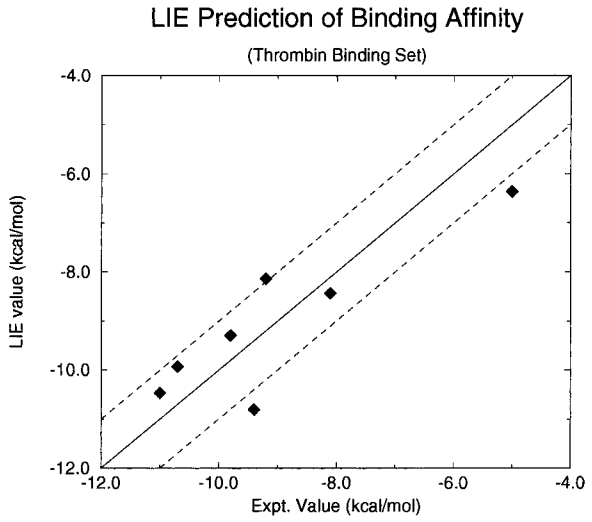

Figure 6. LIE binding energies for the thrombin binding set. The overall RMS error is 0.94 kcal/mol for seven ligands studied here. If LIE results agree perfectly with the experimental values, the data points (represented by diamonds) should be on the diagonal line.

analogues binding to thrombin receptor from a 30 ps MD data collection. The protein and ligand structures are identical to those published by Jorgensen et al.,[13] so interested readers can refer there for detailed molecular structures. As we can see from the figure, the overall agreement with experiments is quite good and the general order ranking is also more or less correct. The overall RMS error is 0.94 kcal/mol and the correlation coefficient is r^2 is 0.759, which are comparable to the best fitting results in explicit solvent LIE model.[13] The fitting parameters are $\alpha = -0.224$, $\beta = 0.0271$, and $\gamma_0 = 0.0566$. The SGB-LIE binding energies for coagulation factor Xa binding set from a 20 ps MD data collection are shown in Figure 7. The protein and ligand structures were kindly supplied by Dr. Daniel Cheney at Bristol Myers Squibb[48] (the detailed structures of the ligands are proprietary and only available upon request). Similarly, the

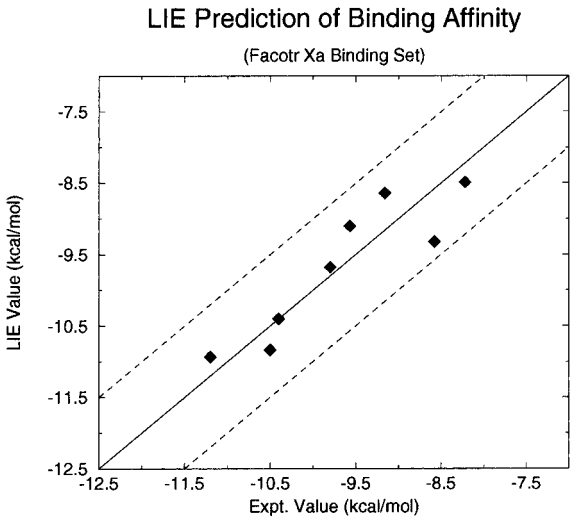

Figure 7. LIE binding energies for the factor Xa binding set. The overall RMS error is 0.41 kcal/mol for eight ligands studied here. If LIE results agree perfectly with the experimental values, the data points (represented by diamonds) should be on the diagonal line.

overall agreement with experiments is quite good. The RMS error for this data sets with eight ligands is only 0.41 kcal/mol and the correlation coefficient is as high as 0.818, indicating a very good agreement with the experimental results. The SGB-LIE parameters for this binding set is $\alpha = 0.144$, $\beta = 0.0935$, and $\gamma_0 = 0.0737$. We think the fitting parameters for different binding sets should not be expected to be the same. In the first few LIE papers, Aqvist et al. suggested that there is a universal parameter $\alpha = 0.168$ and β should be 0.50, but the number of ligands under study were very small (three to seven). Later, Aqvist et al. suggested to use different β values on the basis of the properties of the ligands, such as ionized molecules (0.50), neutral molecules with zero hydroxyl group (0.43), one hydroxyl group (0.37), and two or more hydroxyl groups (0.33).[10] Jorgensen et al. pointed out that a third parameter is necessary to get a good fit for a seven-ligand thrombin binding set,[13] and also found that the parameters are quite different from those of Aqvist et al. We think it is reasonable to assume that the parameters should be system dependent and might be force field dependent also. It should be noted that we used van der Waals energy and SGB cavity energy to estimate the nonpolar or nonelectrostatic contribution of the binding free energy, there is an alternative approach for estimating this portion of the binding free energy. Warshel et al.[21,22] used scaled protein dipoles Langevin dipoles (PDLD/S) to estimate the van der Waals and hydrophobic contributions. Both approaches use reduced models for the explicit water solvent, either by a continuum with a dielectric constant of 80.0 or by Langevin dipoles.

Finally, we compared the CPU timing for our continuum solvation model SGB-LIE with the explicit solvent model LIE. Following previous approaches by Aqvist et al.[8,9] and Jorgensen et al.,[13,14] we built a 20 Å water sphere (SPC water model) around the ligand for the explicit solvation. For the ligand H01 (HEPT), which is used as an example for CPU illustration in the following, this gives a size of total 3604 atoms for the free state and 3298 atoms for the bound state. The solvated systems are first minimized with conjugate gradient minimizer and then equilibrated for 100 ps for both the free and bound states. Figure 8 shows the 200 ps data collection for the time average of the LIE components, van der Waals, and Coulomb energies in the explicit solvent model. In the free state, one can see that there are large fluctuations in the electrostatic energy, and it takes

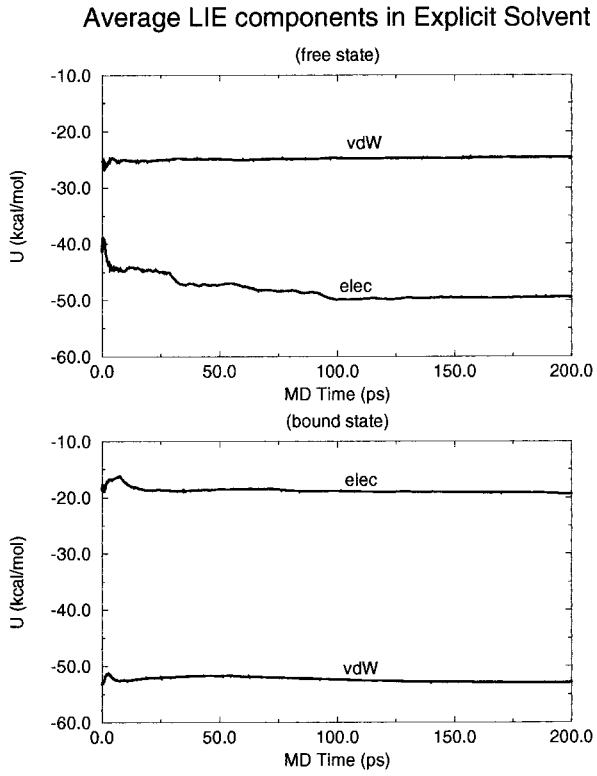

Figure 8. The variation of the average LIE interaction components vs MD time in explicit solvent model. A 20.0 Å SPC water sphere is used as the explicit solvent model. The plot shows a 200 ps MD simulation following a 100 ps MD equilibration for HEPT binding with HIV-1RT. The top one shows the averages for the free state, and the bottom one for the bound state.

about 125 ps before the electrostatic energy converges. The bound state shows relatively smaller fluctuations, since the ligand is confined in the cavity of the active site, and the total number of conformations accessible should be smaller than that in the free state, but it still takes about 100 ps to converge by using the same convergence criterion used in the continuum solvent model (each component changes no more than 0.5 kcal/mol when doubling the data collection time). There is another interesting way checking for the convergence. As we know that for solvation free energy (in free state), the reaction field energy should be half the Coulomb energy.[12] The converged reaction field energy U_{rxn} of HEPT in the continuum free state is −24.3 kcal/mol, and the converged Coulomb energy U_{coul} of HEPT in the explicit free state is −49.1 kcal/mol, showing that there is indeed a factor of 2 between the Coulomb energy and the reaction field energy, which means our electrostatic energies are likely converged. Interestingly, our explicit solvent model's MD convergence times are also surprisingly close to those from Aqvist et al.'s results on HIV protease/inhibitors binding set.[9] A 110 and 130 ps equilibration for the free and bound state is used respectively in Aqvist et al.'s work, and a 125 ps data collection for both the free and bound states is used. Even though the two systems under studying are very different, the MD equilibration time and data collection time are very similar. We think it makes sense since the actual system sizes for both the protein and ligand in the two studies are comparable: both used 20 Å water spheres and distant residue truncations for the protein system. We speculate that, for this type of simulations, about 100 ps MD averaging might be necessary with explicit solvent solvation.

Table 7 shows the CPU timing for explicit and continuum solvent LIE models. The CPU timing is obtained from IBM

TABLE 7: CPU Comparison of the Continuum Solvent LIE Model with the Explicit Solvent LIE Model[a]

model	state	equilibration	data collected	cpu/ps	total cpu
explicit	free	100 ps	125 ps	0.370 h	3.47 d
	bound	100 ps	110 ps	0.336 h	3.08 d
continuum	free	15 ps	30 ps	0.11 min	4.95 min
	bound	15 ps	30 ps	0.291 h	0.546 d

[a] Data is collected for HEPT (H01) binding to HIV-1RT, and the explicit solvent model is based on the solvation of a 20 Å water sphere from the center of the ligand, as used by Aqvist et al.[9] ("min" for minutes, "h" for hours, and "d" for days.)

Power3-375MHZ SP2 clusters. The MD was run with a time step of 2.0 fs in both simulations using multiple time step algorithm RESPA.[31,32] No cutoff for the long-range electrostatic interactions is used in both simulations. As one can see from the table, it takes about 3.47 days for the free state and 3.08 days for the bound state in explicit solvent LIE model (total 6.55 days), but only 4.95 min for the free state and 0.546 days for the bound state in the SGB-LIE model (total 0.549 days), showing a total 12-fold speedup. A 30 ps MD data collection was used in CPU estimation for SGB-LIE, even though we only used 15 ps data collection in previous discussions for the HIV-1RT set, but we did use 20−30 ps for other binding sets and also we tried 30 ps data collection for the HIV-1RT set (see discussions below). In SGB-LIE model, the free state is virtually "free" since it only takes minutes, while the free state is equally expensive as the bound state in the explicit solvent model for obvious reasons, so a factor of 2 in CPU timing is gained immediately. The second major speedup (about a factor of 4−5) is from the fact that a shorter MD run (≈30 ps) seems enough in the continuum solvent model as compared to the much longer MD run in the explicit solvent model (≈125 ps). We think there might be two reasons for this: first, for a given ligand conformation, the average interaction from the solvent is instantaneous in the continuum solvent model, but it takes a significant time to reequilibrate water molecules and converge the interaction energy in the explicit solvent model; second, the conformational change (kinetics) of the ligand in the continuum solvent model should be faster than that in the explicit solvent model due to the lack of collisions with water molecules in the continuum solvent. Another speedup is from each picosecond or each step's CPU time for the bound state. There is about 15% −20% speedup in each ps for the bound states that we have examined. One might be surprised that why there is only a 15−20% speedup in the bound state for each step or each picosecond? The reason is actually very simple. The 20 Å water sphere used in LIE's explicit solvation is far from a complete solvation for the total protein system, since the protein atoms on the boundary are not completely solvated and protein atoms beyond this 20 Å boundary is not solvated at all; on the other hand, the continuum solvent model simulates the whole solvated protein system. For comparison purpose, the same 20 Å water sphere can be regarded as a complete solvation for the ligand in the free state, and the speedup of the SGB model over the explicit solvent model is about 200 (0.370 h/0.11 min)!

We have also run two other simulations for the HIV-1RT binding set for CPU testing purpose (to see how long MD is really needed in addition to the component test we did earlier and how often SGB needs to be updated). One is to run 30 ps MD data collection for all the 20 ligands in the HIV-1RT binding set instead of 15 ps used above (with SGB updated every 10 steps as before), and the other is to run 15 ps MD data collection but with SGB updated every three steps instead of every 10. The overall RMS error for the 30 ps MD is 1.12

kcal/mol with cross-validation Jackknife test error of 1.28 kcal/mol, so they are very close to those from 15 ps MD, which were 1.07 and 1.31 kcal/mol, respectively. The correlation coefficients are 0.756 and 0.678, also similar to those from 15 ps MD data collection. The SGB-LIE fitting parameters are $\alpha = 0.174$, $\beta = 0.335$, and $\gamma = 0.0100$, which are also comparable to those from 15 ps MD and those listed in Table 6. The results from updating SGB every three steps are slightly better, with RMS errors of 1.02 and 1.18 kcal/mol, and the correlation coefficients of 0.801 and 0.730 for the SGB-LIE fitting and cross-validation, respectively. The SGB-LIE fitting parameters are $\alpha = 0.254$, $\beta = 0.434$, and $\gamma = 0.00548$, within the same range as before. The CPU cost for the 30 ps MD is twice that of the 15 ps MD, but we already used the 30 ps MD time for comparison in Table 7; in other words, if we only use 15 ps, the speed up will be even higher. The speed of updating SGB every three steps is about a factor of 2.2 slower than that of updating every 10 steps, but nevertheless it will still be about an order faster than the explicit solvent model. Also, from our fitting results, it seems fine to update SGB every 10 steps for this type of simulation.

4. Conclusion

In this work, we have proposed a new linear interaction energy (LIE) method based on continuum solvent surface generalized Born (SGB) model for ligand−receptor binding affinity prediction. The method SGB-LIE uses three terms in binding free energy formula, van der Waals energy between the ligand and the receptor, electrostatic energy including the Coulomb energy between the ligand and the receptor and reaction field energy between the ligand and the continuum solvent, and cavity energy between the ligand and the continuum solvent. The new method is much faster than the previously proposed LIE methods based on explicit solvents. In the binding sets we tested here, about a 12-fold speedup is achieved. Also, there is no need to add in a Born correction term for ionized systems as is done in explicit solvent models, or to keep the protein system neutral to avoid the Born correction due to the finite size of the solvent sphere.[8,9,13]

The new approach has been applied to three binding sets: HEPT analogues binding to HIV-1 reverse transcriptase (20 ligands), sulfonamide inhibitors binding to human thrombin (seven ligands), and various ligands binding to coagulation factor Xa (eight ligands). The fitting and cross-validation results show that about 1.0 kcal/mol accuracy is achievable for binding sets with as many as 20 ligands. We have also explored various techniques for the underlying LIE conformation space sampling, including molecular dynamics and hybrid Monte Carlo methods, and the final results show that comparable binding energies can be obtained no matter which sampling technique is used. Future studies will be focused on testing even bigger binding sets and also improving the continuum solvation model.

Acknowledgment. We would like to thank Dr. Daniel Cheney for providing us the initial docked structures of the coagulation factor Xa binding set. Part of the calculations were performed at SGI's Supercomputer Center GS2. This work was partially supported by a grant to R.A.F. from the National Institutes of Health (NIH) (GM-52018), and by the NIH, Division of Research Resource (P41-RR06892), and to W.L.J. from the NIH, National Institute of Allergy and Infectious Diseases (AI44616).

References and Notes

(1) Tomioka, N.; Itai, A.; Iitaka, Y. J. *J. Comput. Aided Mol. Design* **1987**, *1*, 197.
(2) Bohm, H. J. *J. Comput. Aided Mol. Design* **1994**, *8*, 243.
(3) Wallqvist, A.; Jernigan, R. L.; Covell, D. G. *Protein Sci.* **1995**, *4*, 1881.
(4) Verkhivker, G.; Appelt, K.; Freer, S. T.; Villafranca, J. E. *Protein Eng.* **1995**, *8*, 677.
(5) Kollman, P. *Chem. Rev.* **1993**, *93*, 2395.
(6) Straatsma, T. P.; McCammon, J. A. *Annu. Rev. Phys. Chem.* **1992**, *43*, 407.
(7) Beveridge, D. L.; Dicapua, F. M. *Annu. Rev. Biophys. Chem.* **1989**, *18*, 431.
(8) Aqvist, J.; Medina, C.; Samuelsson, J. E. *Protein Eng.* **1994**, *7*, 385.
(9) Hanson, T.; Aqvist, J. *Protein Eng.* **1995**, *8*, 1137.
(10) Aqvist, J.; Hanson, T. *J. Phys. Chem.* **1996**, *100*, 9512.
(11) Ljungberg, K. B.; Marelius, J.; Musil, D.; Svensson, P.; Norden, B.; Aqvist, J. *Eur. J. Pharm. Sci.* **2001**, *12*, 441.
(12) Carlson, H. A.; Jorgensen, W. L. *J. Phys. Chem.* **1995**, *99*, 10667.
(13) Jones-Hertzog, D. K.; Jorgensen, W. L. *J. Med. Chem.* **1997**, *40*, 1539.
(14) Rizzo, R. C.; Tirado-Rives, J.; Jorgensen, W. L. *J. Med. Chem.* **2001**, *44*, 145.
(15) Smith, R. H., Jr.; Jorgensen, W. L.; Tirado-Rives, J.; Lamb, M. L.; Janssen, P. A.; Michejda, C. J.; Kroeger Smith, M. B. *J. Med. Chem.* **1998**, *41*, 5272.
(16) Lamb, L.; Tirado-Rives, J.; Jorgensen, W. L. *Bioorg. Med. Chem.* **1999**, *7*, 851.
(17) Pierce, A. C.; Jorgensen, W. L. *J. Med. Chem.* **2001**, *44*, 1043.
(18) Kroeger Smith, M. B.; Lamb, M. L.; Tirado-Rives, J.; Jorgensen, W. L.; Michejda, C. L.; Smith, R. H.; *Protein Eng.* **2000**, *13*, 413−421.
(19) Wall, I. D.; Leach, A. R.; Salt, D. W.; Ford, M. G.; Essex, J. W. *J. Med. Chem.* **1999**, *42*, 5142.
(20) Wang, W.; Wang, J.; Kollman, P. A. *Proteins* **1999**, *34*, 395.
(21) Lee, F. S.; Chu, Z. T.; Bolger, M. B.; Warshel, A. *Protein. Eng.* **1992**, *5*, 215.
(22) Sham, Y.; Chu, Z.; Tao, H.; Warshel, A. *Proteins: Struct., Funct., Genet.* **2000**, *39*, 393.
(23) Lee, F. S.; Chu, Z. T.; Warshel, A. *J. Comput. Chem.* **1993**, *14*, 161.
(24) Warshel, A.; Tao, H.; Fothergill, M.; Chu, Z. T. *Isr. J. Chem.* **1994**, *34*, 253.
(25) Muegge, I.; Tao, H.; Warshel, A. *Protein. Eng.* **1997**, *10*, 1363.
(26) Kitchen, D. B.; Hirata, F.; Westbrook, J. D.; Levy, R. M.; Kofke, D.; Yarmush, M. *J. Comput. Chem.* **1990**, *11*, 1169.
(27) Figueirido, F.; Zhou, R.; Levy, R.; Berne, B. J. *J. Chem. Phys.* **1997**, *106*, 9835.
(28) Ghosh, A.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B* **1998**, *102*, 10983.
(29) Jorgensen, W. L.; Maxwell, D.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225.
(30) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127.
(31) Tuckerman, M.; Berne, B. J.; Martyna, G. J. *J. Chem. Phys.* **1992**, *97*, 1990.
(32) Zhou, R.; Berne, B. J. *J. Chem. Phys.* **1995**, *103*, 9444.
(33) Zhou, R.; Friesner, R. A. **2001**. In preparation.
(34) Wilson, E. K. *Chem. Eng. News* **1996**, *74*, 42.
(35) Cohen, J. *Science* **1997**, *277*, 32.
(36) Tanaka, H.; et al. *J. Med. Chem.* **1992**, *35*, 4713.
(37) Tanaka, H.; et al. *J. Med. Chem.* **1995**, *38*, 2860.
(38) Tanaka, H.; et al. *J. Med. Chem.* **1991**, *34*, 349.
(39) Tanaka, H.; et al., *J. Med. Chem.* **1992**, *35*, 337.
(40) Baba, M. et al. *Antimicrob. Agents Chemother.* **1994**, *38*, 688.
(41) Friedrich, M.; Zhou, R.; Edinger, S.; Friesner, R. A. *J. Phys. Chem.* **1999**, *103*, 3057.
(42) Duane, S.; Kennedy, A. D.; Pendleton, B. J.; Roweth, D. *Phys. Lett. B* **1987**, *195*, 216.
(43) Zhou, R.; Berne, B. J. *J. Chem. Phys.* **1997**, *107*, 9185.
(44) Frantz, D. D.; Freeman, D. L.; Doll, J. D. *J. Chem. Phys.* **1990**, *93*, 2769.
(45) Li, Z.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 6611.
(46) Guarneri, F.; Still, W. C. *J. Comput. Chem.* **1994**, *11*, 1302.
(47) Andricioaei, I.; Straub, J. E. *J. Chem. Phys.* **1997**, *107*, 9117−9124.
(48) Cheney, D. L. Personal communications.