J|A|C|S
ARTICLES

Published on Web 03/10/2009

Path-Integral Method for Predicting Relative Binding Affinities of Protein−Ligand Complexes

Chandrika Mulakala[†] and Yiannis N. Kaznessis*

Department of Chemical Engineering and Materials Science, 151 Amundson Hall, 421 Washington Avenue SE, University of Minnesota, Minneapolis, Minnesota, 55455

Received September 19, 2008; E-mail: yiannis@cems.umn.edu

Abstract: We present a novel approach for computing biomolecular interaction binding affinities based on a simple path integral solution of the Fokker−Planck equation. Computing the free energy of protein−ligand interactions can expedite structure-based drug design. Traditionally, the problem is seen through the lens of statistical thermodynamics. The computations can become, however, prohibitively long for the change in the free energy upon binding to be determined accurately. In this work, we present a different approach based on a stochastic kinetic formalism. Inspired by Feynman's path integral formulation, we extend the theory to classical interacting systems. The ligand is modeled as a Brownian particle subjected to the effective nonbonding interaction potential of the receptor. This allows the calculation of the relative binding affinities of interacting biomolecules in water to be computed as a function of the ligand's diffusivity and the curvature of the potential surface in the vicinity of the binding minimum. The calculation is thus exceedingly rapid. In test cases, the correlation coefficient between actual and computed free energies is >0.93 for accurate data sets.

Introduction

The accumulation of atomic-resolution information of protein−ligand interactions available through public databases of X-ray crystallographic and NMR structures[1] has paved the way for the development of several theoretical methods for binding free energy prediction[2] based on atomistic representations of the protein−ligand complex. Typically, the theoretical foundation is based on statistical thermodynamics. For an interaction between a protein and a ligand, E and I, interacting in solution to form a complex E*I, we can write the biomolecular reaction

$$[\mathrm{E}]_{\mathrm{aq}} + [\mathrm{I}]_{\mathrm{aq}} \underset{k_{\mathrm{off}}}{\overset{k_{\mathrm{on}}}{\rightleftharpoons}} [\mathrm{E^*I}]_{\mathrm{aq}} \tag{1}$$

At equilibrium, we can write for the standard free energy of association

$$\Delta G^{\circ} = \mu^{\circ}_{\mathrm{E*I}} - (\mu^{\circ}_{\mathrm{E}} + \mu^{\circ}_{\mathrm{I}}) = -RT\ln(1/K_{\mathrm{i}}) \tag{2}$$

where $\mu°_{\mathrm{E*I}}$, $\mu°_{\mathrm{E}}$, and $\mu°_{\mathrm{I}}$ are the standard chemical potentials of the complex and the individual species, respectively, R is the ideal gas constant, T is the temperature, and $1/K_{\mathrm{i}} = k_{\mathrm{on}}/k_{\mathrm{off}}$ is the binding equilibrium constant.

These macroscopic thermodynamic properties connect to microscopic properties determined by atomistic computer simulations through the classical statistical thermodynamics relationship

$$\Delta G^{\circ} = \Delta A^{\circ} + P\Delta V^{\circ} = -RT\ln[\{Q_{\mathrm{E*I}}/(N_{\mathrm{Avo}}Q_{\mathrm{W}})\}/ \{(Q_{\mathrm{E}}/(N_{\mathrm{Avo}}Q_{\mathrm{W}}))(Q_{\mathrm{I}}/(N_{\mathrm{Avo}}Q_{\mathrm{W}}))\}] + P\Delta V^{\circ} \tag{3}$$

where $Q_{\mathrm{E*I}}$, Q_{E}, Q_{I}, and Q_{W} are the molecular, canonical ensemble partition functions of the complex, the individual species, and the solvent, respectively. In principle, the partition functions enumerate all the possible microscopic states of the molecules. In practice, the direct calculation of the partition function for as complex a system as a solvated protein is theoretically and computationally unfeasible, because of the configurational integral. For one of the interacting species, say inhibitor *I*, this integral is

$$Z_{\mathrm{I}}^{\mathrm{int}} = \int ... \int \exp\{-E(x_{\mathrm{I}}^{N}, x_{\mathrm{W}}^{M})/k_{\mathrm{B}}T\}\mathrm{d}x_{\mathrm{I}}^{N}\mathrm{d}x_{\mathrm{W}}^{M} \tag{4}$$

where x_{I}^{N} and x_{W}^{M} are the dimensions of the configurational space available to the N molecules of species I and the M molecules of solvent W. $E(x_{\mathrm{I}}^{N}, x_{\mathrm{W}}^{M})$ is the potential energy of interaction between I and W.

There have been ingenious efforts to approximate the calculation, decomposing the free energy into numerous components that can be tractably computed.[3−14] Nonetheless,

[†] Current Address: Department of Electrical and Computer Engineering, EECS Building, 200 Union Street SE, University of Minnesota, Minneapolis, MN 55455.

(1) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(2) Pohorille, A., Chipot, C., Eds. *Free energy calculations - Theory and applications in chemistry and biology*; Springer: Heildelberg, 2007.

(3) Gräter, F.; Schwarzl, S. M.; Dejaegere, A.; Fischer, S.; Smith, J. C. *J. Phys. Chem. B* **2005**, *109*, 10474–10483.

(4) Oostenbrink, C.; van Gunsteren, W. F. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6750–6754.

(5) Raha, K.; Merz, K. M., Jr. *J. Med. Chem.* **2005**, *48*, 4558–4575.

(6) Huang, N.; Kalyanaraman, C.; Bernacki, K.; Jacobson, M. P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5166–5177.

(7) Laurie, A. T.; Jackson, R. M. *Curr. Protein Pept. Sci.* **2006**, *7*, 395–406.

10.1021/ja807460s CCC: $40.75 © 2009 American Chemical Society

difficulties persist, mainly because the error in the calculations of free energy components is larger than the actual, absolute value of the binding strength. Therefore, although theoretically on firm ground, the statistical thermodynamics approach to use atomistic models of biological molecules to predict the free energy of binding can be fruitful only with the injection of empirically derived corrections or severe simplifying assumptions. We propose the following alternative approach.

Considering the phase orbit of a classical mechanical system in phase space and integrating out the solvent degrees of freedom, the path will resemble the motion of a Brownian particle described by a Langevin equation:

$$\frac{\mathrm{d}\underline{x}}{\mathrm{d}t} = -\frac{D}{k_{\mathrm{B}}T}\frac{\partial U(\underline{x})}{\partial \underline{x}} + W(t) \tag{5}$$

where $W(t)$ is Gaussian noise with average $\langle W(t)\rangle = 0$ and correlation $\langle W(t)W(t')\rangle = 2D\delta(t - t')$, D is the diffusion of the system in phase space, and k_{B} and T are the Boltzmann factor and temperature, respectively.

An equivalent Fokker–Planck equation is

$$\frac{\mathrm{d}P(\underline{x}, t)}{\mathrm{d}t} = D\frac{\partial}{\partial \underline{x}}\left(\frac{1}{k_{\mathrm{B}}T}\frac{\partial U(\underline{x})}{\partial \underline{x}}P(\underline{x}, t)\right) + D\frac{\partial^2 P(\underline{x}, t)}{\partial \underline{x}^2} \tag{6}$$

The solution can be written as a path integral[15−18]

$$P(\underline{x}_f, t_f|\underline{x}_i, t_i) = e^{-(U(\underline{x}_f)-U(\underline{x}_i))/k_{\mathrm{B}}T}\int \mathscr{D}\underline{x}[t]e^{-S_{\mathrm{eff}}[\underline{x}]} \tag{7}$$

where,

$$S_{\mathrm{eff}}[\underline{x}] = \int_{t_i}^{t_f} d\tau[\dot{\underline{x}}^2(\tau)/4D + V_{\mathrm{eff}}(\underline{x})] \tag{8}$$

The effective potential is

$$V_{\mathrm{eff}}(\underline{x}) = \frac{D}{2}\left(\frac{1}{k_{\mathrm{B}}T}\frac{\partial U(\underline{x})}{\partial \underline{x}}\right)^2 - \frac{D}{k_{\mathrm{B}}T}\frac{\partial^2 U(\underline{x})}{\partial \underline{x}^2} \tag{9}$$

The conditional probability solution for remaining in the same state space position for small time intervals is

$$\lim_{t\to 0} P(\underline{x}, \underline{x};t) = \frac{1}{\sqrt[n]{4\pi Dt}}\exp(-V_{\mathrm{eff}}(\underline{x})t) \tag{10}$$

Here t is very short, of the order of the average duration of solvent collisions, and n is the number of degrees of freedom.

In a lucid analysis of the Langevin equation, de Grooth[19] surmises that the friction coefficient $\gamma = k_{\mathrm{B}}T/D = 2m_sf$, where m_s is the mass of the solvent molecules and f is the number of collisions per second, giving the average time per collision as $2m_sD/k_{\mathrm{B}}T$. Therefore,

$$\lim_{t\to 0} P(\underline{x}, \underline{x};t) = \sqrt[\mathrm{n}]{\frac{k_{\mathrm{B}}T}{8\pi m_s D^2}}\exp\left(-\frac{2m_sDV_{\mathrm{eff}}(\underline{x})}{k_{\mathrm{B}}T}\right) \tag{11}$$

If we consider the reversible reaction between the two areas of phase space representing the bound and unbound states of the ligand given by

$$[I_{\mathrm{bound}}]_{\mathrm{aq}} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} [I]_{\mathrm{aq}} \tag{12}$$

where $[I_{\mathrm{bound}}]_{\mathrm{aq}} = [\mathrm{E^*I}]_{\mathrm{aq}}$, the transitional probability to stay in the bound conformation (denoted by $\underline{x}_{\mathrm{B}}$) is[20]

$$\lim_{t\to 0} P(\underline{x}_{\mathrm{B}}, \underline{x}_{\mathrm{B}};t) = \frac{k_{-1}}{k_1 + k_{-1}} \tag{13}$$

Combining equations 11 and 13, we get

$$\sqrt[\mathrm{n}]{\frac{k_{\mathrm{B}}T}{8\pi m_s D^2}}\exp\left(-\frac{2m_sDV_{\mathrm{eff}}(\underline{x})}{k_{\mathrm{B}}T}\right) = \frac{k_{-1}}{k_1 + k_{-1}} \tag{14}$$

For equations 1 and 12 to be equivalent, since $[I_{\mathrm{bound}}]_{\mathrm{aq}} = [\mathrm{E^*I}]_{\mathrm{aq}}$, $k_{\mathrm{on}} = k_{-1}/[\mathrm{E}]$ and $k_{\mathrm{off}} = k_1$. Substituting these and for V_{eff} from eq 9, at the minimum energy bound conformation of [E*I],

$$\ln\left(1 + \frac{K_{\mathrm{i}}}{[\mathrm{E}]}\right) = -\left(\frac{2m_sD^2}{(k_{\mathrm{B}}T)^2}\frac{\partial^2 U(\underline{x})}{\partial \underline{x}^2} + \ln\left(\sqrt[\mathrm{n}]{\frac{k_{\mathrm{B}}T}{8\pi m_s D^2}}\right)\right) \tag{15}$$

where K_{i} is the equilibrium dissociation constant for eq 1 and [E] is the free enzyme concentration (see Supporting Information for a derivation).

Equations 14 and 15 provide an elegant means for computing the free energy of biomolecular interactions given the bound state structure. Equation 15 has only two variables: the diffusion coefficient of the ligand, D, and the second derivatives of the potential for the bound complex. The latter can be computed quickly given the bound state, as described later.

Diffusion coefficients for organic molecules can also be estimated to high accuracy by semiempirical derivatives of the Stokes–Einstein relation which typically give scaling functions $D \approx V^{-0.6}$, where V is the molar volume of the solute.[21]

Even though, in theory, the right-hand side (RHS) of eq 15 can be exactly determined, experimental errors would give rise to empirical coefficients for the two terms in RHS. In practice, therefore, these variables in eq 15 must be trained with a set of protein–ligand complexes with known binding free energies.

Here we present a systematic study of the application of eq 15 toward the prediction of binding affinities of three different enzyme–inhibitor systems: bovine trypsin, β-secretase, and aldose reductase. We restrict our study to systems for which experimental inhibitor affinity measures (K_{i}, IC*50*) as well as X-ray crystal structures for the bound complexes are available. We find that eq 15 is not only a convenient and accurate means for predicting binding affinities but it also provides useful

[8] Foloppe, N.; Hubbard, R. *Curr. Med. Chem.* **2006**, *13*, 3583–3608.
[9] Gilson, M. K.; Zhou, H. X. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.
[10] Mobley, D. L.; Graves, A. P.; Chodera, J. D.; McReynolds, A. C.; Shoichet, B. K.; Dill, K. A. *J. Mol. Biol.* **2007**, *371*, 1118–1134.
[11] Ruvinsky, A. M. *J. Comput. Aided Mol. Des.* **2007**, *21*, 361–370.
[12] Sega, M.; Faccioli, P.; Pederiva, F.; Garberoglio, G.; Orland, H. *Phys. Rev. Lett.* **2007**, *99*, 118102–118104.
[13] Ghosh, A.; Elber, R.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 10394–10398.
[14] Zeevaart, J. G.; Wang, L. G.; Thakur, V. V.; Leung, C. S.; Tirado-Rives, J.; Bailey, C. M.; Domaoal, R. A.; Anderson, K. S.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2008**, *130*, 9492–9499.
[15] Onsager, L.; Machlup, S. *Phys. Rev.* **1953**, *91*, 1505–1512.
[16] Feynman, R. P.; Hibbs, A. R. *Quantum mechanics and path integrals*; McGraw-Hill: Maidenhead, 1965.
[17] Wiegel, F. W. *Physica* **1967**, *37*, 105–113.
[18] Wiegel, F. W. *Physica* **1967**, *33*, 734–736.
[19] de Grooth, B. G. *Am. J. Phys.* **1999**, *67*, 1248–1252.
[20] McQuarrie, D. A. *J. Appl. Prob.* **1967**, *4*, 413–478.
[21] Wilke, C. R.; Chang, P. *A. I. Ch. E. J.* **1955**, *1*, 264–270.

insights on the effect of the oft-ignored assay conditions on measured binding affinities.

Computational Methods

Diffusivities for organic molecules at infinite dilutions can be accurately determined by semiempirical derivatives of the Stokes−Einstein relationship, of which the following equation by Wilke-Chang is perhaps the most popular,[21]

$$D_{\mathrm{WC}} = \frac{7.4 \times 10^{-12}\sqrt{\chi M_{\mathrm{W}}}T}{\eta V^{0.6}} \tag{16}$$

where χ indicates the degree of solvent association, M_{w} is the molar mass of the solvent (18.015 g mol^{-1}), η is solvent viscosity (mPa s) at temperature T (K), and V is the Le Bas molar volume[22] of the solute at its normal boiling point. It has been shown previously[23] that the van der Waals volume of the molecule in Å^3 (V_{vdw}) estimated by molecular modeling software can approximate the Le Bas volume. V_{vdw} was therefore calculated in MOE[24] using a grid approximation with a spacing of 0.75 Å.

The first and second terms on the RHS of eq 15 (denoted as TermA and TermB, respectively) were computed using D_{WC} for diffusivity D. These computed terms were used to train the linear regression fit, giving an equation of the form:

$$\ln[K_{\mathrm{i}}'] = \mathrm{A}^*\mathrm{TermA} + \mathrm{B}^*\mathrm{TermB} + \mathrm{I} \tag{17}$$

where A and B are empirical constants and I is the intercept. For the trypsin data set $ln[K_{\mathrm{i}}'] = ln[1 + K_{\mathrm{i}}/[E]]$, where the binding constant, K_m, and the total enzyme concentration, $[E_{\mathrm{Total}}]$ were known. The binding affinity data were approximated as $\ln[1{+}K_{\mathrm{i}}]$ or $\ln[\mathrm{IC}_{50}]$ for the other two sets.

Since χ is a dimensionless empirical parameter, different values have been proposed for χ in the literature for different types of molecules: 2.6 by Wilke and Chang,[21] 2.9 for organic electrolytes,[25] 2.26 for nonelectrolytes,[26] and 1.61 for aromatics.[27] In this study, we used $\chi = 2.6$ and introduced a new empirical parameter φ, such that $D = \varphi\ {}^*D_{\mathrm{WC}}$, to correct for the use of V_{vdw} instead of Le Bas volume, as well as for the presence of other solutes in the assay buffer. A second normalizing term, λ, was introduced to eliminate the intercept.

The modified eq 15 in three dimensions is therefore,

$$\ln\left(1 + \frac{K_{\mathrm{i}}}{[\mathrm{E}]}\right) = -\kappa\left(\frac{2m_s\phi^2D_{\mathrm{WC}}^2}{(k_{\mathrm{B}}T)^2}\frac{\partial^2U(\underline{x})}{\partial\underline{x}^2} + \ln\left(\lambda^3\sqrt[3]{\frac{k_{\mathrm{B}}T}{8\pi m_s\phi^2D_{\mathrm{WC}}^2}}\right)\right) \tag{18}$$

Note that parameters φ and λ are not necessary to obtain a useful equation for predicting activities. However, determination of φ helps to compare the effective diffusivities of the ligands between data sets, while parameter λ nondimensionalizes the probability of eq 10.

For the first derivative of the potential to be zero in V_{eff} (eq 9), the solvated enzyme/ligand complex has to be proximal to its lowest energy conformation. Protein/ligand complexes were therefore minimized prior to computation of the second derivatives. The ligand/receptor complexes were minimized in MOE[24] to a rms gradient of 0.001 using the MMFF94 force field.[28]The rms gradient is the product of norm of the potential gradient and the square root of the number of unfixed atoms. Nonbonded interactions were evaluated without any cut-offs. During minimization, the solvent was implicitly modeled through the Generalized-Born model as implemented in MOE. Only residues of the receptor within 5 Å of the ligand were selected for the minimization since we were only interested in minor side-chain modifications in the vicinity of the ligand.

If we treat the inhibitor as a rigid body at the potential minumum, $\partial^2U(\underline{x})/\partial\underline{x}^2$, of eq 15 is the trace of the resultant three-dimensional Hessian matrix ($k = k_{xx} + k_{yy} + k_{zz}$). The components of the Hessian matrix were evaluated from a finite difference of the forces about the potential minimum as follows:

$$k_{ij} = \frac{\partial^2U}{\partial i\partial j} = \frac{\partial F_i}{\partial j} = \frac{F_{i+\Delta i} - F_{i-\Delta i}}{2\Delta j} \tag{19}$$

with $\Delta i = \Delta j = 1 \times 10^{-6}$ Å, where i and j refer to the three Cartesian dimensions x,y and z, F_i is the force at position i. Forces were evaluated using the Potential[] function in MOE. The trace of the Hessian is therefore $k = k_{xx} + k_{yy} + k_{zz}$.

Statistical analysis and linear regression was carried out using JMP.[29] The leave-one-out cross validation coefficient, $R^2{}_{\mathrm{cv}}$, and the root-mean-square error of cross validation (s_{cv}) were computed using MATLAB.[30]

Results

TermA and TermB of eq 17 were computed for the entire data set (Table 1). The linear regression fit for the entire data set is presented in figure 1. While the R^2 of the fit is not very significant (0.51), a close examination of Figure 1 presents some remarkable trends. Significant clustering is observed for points belonging to the same protein, with either systematic over- or under prediction within data-points of those subsets. We assumed that these differences were due to either the specific details of the assay buffers, or the chemical characteristics of ligands themselves, both of which would affect the ligands' diffusivities. The Wilke-Chang equation (eq 16) evaluates diffusivities for molecules at infinite dilution, and systematic deviations from the calculated diffusivities are expected that are dependent on the assay buffers. We therefore decided that each subset had to be trained separately for better fit. Also intriguing was the presence of two clusters in the β-secretase data set, one of which was significantly underpredicted (Figure 1) and belonged entirely to data from the work of the same group presented in two different papers.[31,32] A comparison of the assays between this subset and the rest of the β-secretase complexes revealed that these assays had 10% dimethylsulfoxide (DMSO) in the assay buffer. DMSO is a highly viscous solvent and is bound to have significant effect on the diffusivity of the ligand molecules. Aminabhavi and Gopalakrishna[33] report a

[22] Le Bas, G. In *The Properties of Gases and Liquids, Monograph*; Longmans, Green and Co.: New York, 1915.

[23] La-Scalea, M. A.; Menezes, C. M. S.; Ferreira, E. I. *J. Mol. Struct.-Theochem* **2005**, *130*, 111–120.

[24] *MOE: The Molecular Operating Environment*, 2006.08 ed.; Chemical Computing Group: Montreal, Canada, 2005.

[25] van der Wielen, L. A. M.; Zomerdijk, M.; Houwers, J.; Luyben, K. C. A. M. *Chem.−Eng. J.* **1997**, *66*, 111–121.

[26] Hayduk, W.; Laudie, H. *A. I. Ch. E. J.* **1974**, *20*, 611–615.

[27] Niesner, R.; Heintz, A. *J. Chem. Eng. Data* **2000**, *45*, 1121–1124.

[28] Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 490–519.

[29] *JMP Statistics Software*, 4.0.4 ed.; SAS Institute: Cary, NC, 1989.

[30] *MATLAB*, 7.2.0.294 (R2006a) ed.; The MathWorks, Inc.: Natick, MA, 2006.

[31] Congreve, M.; Aharony, D.; Albert, J.; Callaghan, O.; Campbell, J.; Carr, R. A. E.; Chessari, G.; Cowan, S.; Edwards, P. D.; Frederickson, M.; McMenamin, R.; Murray, C. W.; Patel, S.; Wallis, N. *J. Med. Chem.* **2007**, *50*, 1124–1132.

[32] Murray, C. W.; Callaghan, O.; Chessari, G.; Cleasby, A.; Congreve, M.; Frederickson, M.; Hartshorn, M. J.; McMenamin, R.; Patel, S.; Wallis, N. *J. Med. Chem.* **2007**, *50*, 1116–1123.

[33] Aminabhavi, T. M.; Gopalakrishna, B. *J. Chem. Eng. Data* **1995**, *40*, 856–861.

Table 1. Dataset of Structures Used for Training the Statistical Linear Regression Fit

PDB	K_i/IC$_{50}$[a] (nM)	ln [K_i'][b]	T (K)	μ (mPa s)	$D_{WC}\times 10^9$ (m^2/s)	k[c] (J/m^2)	TermA	TermB	R_{full}[d]	R_{subset}[d]
Subset 1: Bovine trypsin										
1ghz	*16000*	8.764	298	0.891	0.640	202.6	0.294	76.418	1.780	0.859
1gi2	*3600*	7.273	298	0.891	0.635	202.5	0.289	76.443	0.302	−0.665
1gi6	*1700*	6.523	298	0.891	0.627	218.5	0.304	76.481	−0.064	−0.129
1o2k	*120*	3.882	298	0.891	0.533	248.1	0.249	76.968	−1.784	−0.988
1o2q	*21*	2.189	298	0.891	0.509	294.0	0.270	77.105	−2.636	−0.043
1o2x	*1400*	6.329	298	0.891	0.507	233.0	0.212	77.116	0.609	0.768
1o30	*170*	4.227	298	0.891	0.466	256.7	0.197	77.373	−0.788	0.411
1o33	*1800*	6.580	298	0.891	0.642	219.2	0.320	76.408	−0.014	0.144
1o36	*1100*	6.088	298	0.891	0.454	187.1	0.136	77.449	0.372	−0.560
1o38	*150*	4.103	298	0.891	0.491	268.6	0.229	77.215	−0.977	0.571
1o3d	*74*	3.405	298	0.891	0.531	270.7	0.270	76.978	−1.883	−0.108
1o3g	*11*	1.601	298	0.891	0.525	296.1	0.289	77.014	−3.256	−0.443
Subset 2a: β-secretase (with 10% v/v DMSO)										
2ohk	2000000	14.509	298	0.891	0.842	105.1	0.264	75.595	3.984	−0.190
2ohl	2000000	14.509	298	0.891	0.847	168.7	0.428	75.577	6.594	0.175
2ohm	310000	12.644	298	0.891	0.692	102.5	0.174	76.186	2.846	−0.293
2ohn	500000	13.122	298	0.891	0.696	120.1	0.206	76.167	3.778	0.203
2ohp	94000	11.451	298	0.891	0.620	161.9	0.220	76.514	3.628	−0.257
2ohq	25000	10.127	298	0.891	0.532	245.9	0.246	76.976	4.435	0.045
2ohr	100000	11.513	298	0.891	0.571	121.3	0.140	76.763	3.306	0.436
2ohs	40000	10.597	298	0.891	0.541	148.5	0.154	76.924	3.212	0.098
2oht	9100	9.116	298	0.891	0.541	141.0	0.146	76.924	1.606	−1.402
2ohu	4200	8.343	298	0.891	0.454	293.5	0.214	77.451	3.892	−0.218
2f3e	156	5.050	298	0.891	0.342	369.5	0.153	78.296	2.750	−0.815
1tqf	1400	7.244	298	0.891	0.389	335.6	0.179	77.914	3.955	0.159
Subset 2b: β-secretase (no DMSO)										
1xs7	*1.6*	0.956	310	0.692	0.335	238.0	0.087	78.423	−1.947	−0.124
1ym2	10	2.303	298	0.891	0.333	301.1	0.118	78.381	−0.256	1.643
1ym4	39	3.664	298	0.891	0.361	403.8	0.186	78.140	1.314	2.566
2b8v	98	4.585	298	0.891	0.397	284.2	0.158	77.853	0.725	0.140
2fdp	26	3.258	298	0.891	0.374	270.1	0.134	78.029	−0.349	−0.219
2f3f	190	5.247	298	0.891	0.401	265.6	0.151	77.823	1.155	0.330
2g94	*0.3*	0.262	310	0.692	0.362	395.1	0.169	78.189	−2.175	−0.816
2iqg	5	1.609	310	0.692	0.385	271.2	0.131	78.003	−2.132	−2.166
2irz	12	2.485	310	0.692	0.416	382.1	0.216	77.770	−0.741	−1.208
2is0	200	5.298	310	0.692	0.416	400.2	0.226	77.770	2.239	1.875
2oah	11	2.398	298	0.891	0.409	331.0	0.196	77.765	−1.185	−1.885
2ph6	27	3.296	310	0.692	0.418	381.1	0.218	77.754	0.038	−0.498
2q11[e]	2775	7.928	298	0.891	0.446	213.9	0.150	77.506	2.647	0.082
2q15[e]	33.9	3.523	298	0.891	0.394	312.8	0.172	77.874	−0.039	−0.376
2vie	33	3.497	298	0.891	0.373	299.5	0.148	78.036	0.143	0.453
2vj6	13	2.565	298	0.891	0.373	326.3	0.161	78.038	−0.571	−0.123
2vj7	40	3.689	298	0.891	0.388	264.2	0.141	77.917	−0.215	−0.623
2vj9	180	5.193	298	0.891	0.402	311.9	0.178	77.817	1.519	0.931
2p4j	*1.1*	0.742	310	0.692	0.346	429.3	0.168	78.320	−1.222	0.843
Subset 3: Aldose reductase										
1el3	108	4.682	298	0.891	0.498	138.5	0.122	77.170	−2.308	1.894
1iei	44	3.784	298	0.891	0.528	344.1	0.340	76.996	−0.312	0.269
1mar	60	4.094	298	0.891	0.520	442.9	0.423	77.044	1.536	0.772
1pwl	73	4.290	298	0.891	0.524	387.1	0.375	77.022	0.874	0.881
1pwm	9	2.197	298	0.891	0.668	160.9	0.254	76.292	−5.905	−4.204
1t41	11	2.398	298	0.891	0.542	327.9	0.340	76.919	−1.967	−1.431
1us0	30	3.401	298	0.891	0.540	328.8	0.339	76.932	−0.948	−0.378
1x97	570	6.346	298	0.891	0.666	229.7	0.360	76.302	0.006	−0.019
1z3n	5	1.609	298	0.891	0.549	287.7	0.307	76.877	−3.453	−2.393
1z8a	0.5	−0.693	298	0.891	0.609	241.4	0.317	76.569	−6.747	−5.960
1z89	0.84	−0.174	298	0.891	0.608	367.0	0.480	76.571	−3.555	−5.439
2fzd	13	2.565	298	0.891	0.559	303.4	0.336	76.824	−2.231	−1.655
2ine	*96000*	11.472	298	0.891	0.899	100.7	0.288	75.399	0.616	1.405
2inz	*3500*	8.161	298	0.891	0.873	100.4	0.271	75.487	−2.651	−1.548
2iq0	*68600*	11.136	298	0.891	0.940	68.0	0.213	75.267	−1.441	0.528
2is7	*4400*	8.390	298	0.891	0.797	90.8	0.204	75.761	−2.491	−0.188
2i16	30	3.401	298	0.891	0.538	329.8	0.337	76.942	−0.932	−0.336
2i17	30	3.401	298	0.891	0.538	328.1	0.336	76.942	−0.960	−0.336
2ikg	530	6.273	298	0.891	0.646	306.3	0.452	76.393	1.762	0.275
2ikh	4100	8.319	298	0.891	0.712	260.6	0.467	76.100	2.971	1.121
2iqd	*25500*	10.146	298	0.891	0.938	67.7	0.210	75.274	−2.438	−0.432
2nvc	550	6.310	303	0.798	0.565	261.5	0.285	76.819	0.679	2.071
2nvd	140	4.942	303	0.798	0.623	71.3	0.095	76.524	−4.886	−0.500

[a] K_is in italics. [b] Where ln[K_i'] is ln[1 + K_i/[E]] (see Supporting Information) for subset1 where [S] = K_m and [E] = 10nM, and ln[1+K_i] or ln[IC$_{50}$] elsewhere. [c] $k = k_{xx}+k_{yy}+k_{zz}$ is the trace of the Hessian matrix. [d] R is the residual (the difference between actual and predicted ln[K_i']). [e] K_is converted to IC$_{50}$s from K_m = 24 μM and [S] = 50 μM.

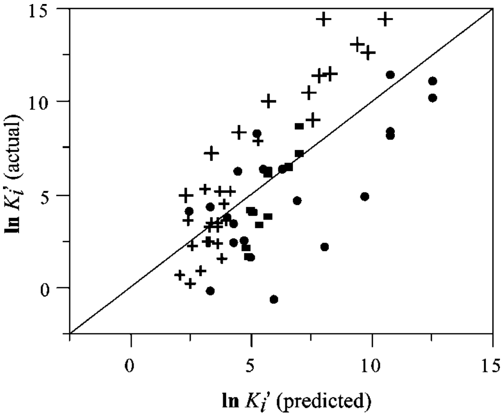

Figure 1. Linear regression fit for the entire data set: ■ (subset 1), † (subset 2a), + (subset 2b), and ● (subset 3).

viscoscity (mPa s) increase to 1.787 from 0.891 as the mole fraction of water changes from 1 to 0.9. A 10% DMSO buffer, therefore increases the buffer viscoscity by ~28%, sufficient to cause the relative shift observed in Figure 1. It is also quite likely that other effects of DMSO, such as altered solubility and hydrophobicity, play an important role toward the activity as well. These data points were therefore separated into two separate subsets for training the linear regression fit.

The statistics of the linear regression fit for the four different subsets belonging to bovine trypsin, β-secretase (with and without DMSO) and aldose reductase are presented in Table 2.

Bovine Trypsin. Experimental evaluations of the inhibitor affinity measurements are very sensitive to experimental conditions with typical coefficients of variation of ~20% in the reported analytical measurements. The best scenario therefore involves data obtained by the same laboratory and the use of the same assay. The trypsin data set, therefore, was of very high quality, since all of the data points used in that set were obtained by the same team. Katz et al.[34,35] conducted a high-throughput study involving the determination of over 300 high resolution crystal structures of trypsin bound with small molecules inhibitors (2-(2-phenol-indoles and 2-(2-phenol)-benzamidazoles) over a wide range of pH. Scanning through the available structures led to a data set of 12 inhibitors with their inhibition constants (K_is). This reduced data set was due to the fact that all the inhibitors studied here exhibit a parabolic pH dependence with a minimum in K_i and therefore only structures determined at pHs close to K_i(min) were used, to ensure that the selected crystal structures captured the protonation state of the inhibitors without any ambiguity. At pH's away from K_i(min) a mixture of protonation states is expected for the inhibitors thereby complicating the picture. This high resolution data set involving crystal structures for all the inhibitors ensures that any errors due to *in silico* mutations and the subsequent molecular mechanics minimization step are avoided. The linear regression fit for this data set is shown in figure 2.

While there are no outliers in this data set, the hydroxyl groups of structures 1o36 and 1o3g had to be protonated for a better fit. This was not unreasonable considering that the pK_a's of the hydroxyl groups for the free molecules are predicted to be close to the pH of the assays. The activities are predicted to very high accuracy ($R^2 = 0.93$, $\sigma = 0.62$, Table 2) in the prediction of ln[K_i']. A $\sigma = 0.62$ implies that the K_is are predicted to the correct order of magnitude, such resolution is rarely seen in computational free energy prediction.

***β*-Secretase.** The β-secretase data set is derived from the work of several independent laboratories each with differing assay conditions, substrates, substrate- and enzyme concentrations and measures of affinity (K_i and IC$_{50}$s) (Table 1). The LHS of eq 18 has the term 1 + K_i/[E] (= IC$_{50}$/[E*I]) (see Supporting Information). The concentration of the enzyme used in these assays was not available for most of the data-points and was therefore approximated by ln[1 + K_i] or ln[IC$_{50}$] for consistency. This approximation is expected to increase the noise in the data set to above that of experimental error. The linear regression fit for the two subsets of β-secretase are presented in Figures 3 and 4. 1tqf and 2f3e were included in data set 1 based on the regression fit. It was not possible for us to determine if that was reasonable, as details of their assays[36,37] are not readily available. Congreve et al. point that the inhibition data for the inhibitor of 2oht should be treated with caution as they exhibited a high hill slope.[31] Interestingly, 2oht is identified as an outlier in the linear regression, and in fact, its inclusion in the fit significantly reduces the accuracy of the fit ($R^2 = 0.95$, $\sigma = 0.5488$, $F = 63.43$), to below the accuracy obtained by including TermB alone in the fit ($R^2 = 0.95$, $\sigma = 0.5213$, $F = 140.3626$). This underscores the importance of accuracy in the data set in training the regression fit. For this data set, TermB dominates in the fit, with TermA offering further resolution to the prediction of activities. In fact the agreement between predicted and actual IC$_{50}$'s is remarkable, a $\sigma = 0.2913$ implying that the predicted IC$_{50}$s are within ~35% of experimental K_i's, with experimental error itself typically of the order of 20%. The domination of TermB in the prediction formula is due to the suppressed values of the diffusivities obtained (Table 2), which appears to be the effect of DMSO in the assay. Since, in TermA the trace of the Hessian matrix is multiplied by D^2, its value is very sensitive to the diffusion coefficient. Significantly lower values of D therefore increase the significance of TermB toward the inhibitory activity.

The second data set has five outliers, 1ym1, 1ym2, 2iso, 2irz, and 2iqg. A sixth outlier, the ligand of 2oah, was assayed without separation of its stereoisomers, therefore justifying its exclusion from the fit. Due to the significant reduction in the accuracy of prediction with the inclusion of 2oht in the previous data set, we chose to exclude 1ym1, 1ym2, 2iqg, 2iso, and 2irz from the fit, even though we are aware that the exclusion of these points artificially improves the statistics of the fit. Supporting our choice is the fact that the four of the outliers are from the work of two different papers,[38,39] and three of these,

[34] Katz, B. A.; Elrod, K.; Luong, C.; Rice, M. J.; Mackman, R. L.; Sprengeler, P. A.; Spencer, J.; Hataye, J.; Janc, J.; Link, J.; Litvak, J.; Rai, R.; Rice, K.; Sideris, S.; Verner, E.; Young, W. *J. Mol. Biol.* **2001**, *307*, 1451–1486.

[35] Katz, B. A.; Elrod, K.; Verner, E.; Mackman, R. L.; Luong, C.; Shrader, W. D.; Sendzik, M.; Spencer, J. R.; Sprengeler, P. A.; Kolesnikov, A.; Tai, V. W.; Hui, H. C.; Breitenbucher, J. G.; Allen, D.; Janc, J. W. *J. Mol. Biol.* **2003**, *329*, 93–120.

[36] Coburn, C. A.; Stachel, S. J.; Li, Y. M.; Rush, D. M.; Steele, T. G.; Chen-Dodson, E.; Holloway, M. K.; Xu, M.; Huang, Q.; Lai, M. T.; DiMuzio, J.; Crouthamel, M. C.; Shi, X. P.; Sardana, V.; Chen, Z. G.; Munshi, S.; Kuo, L.; Makara, G. M.; Annis, D. A.; Tadikonda, P. K.; Nash, H. M.; Vacca, J. P.; Wang, T. *J. Med. Chem.* **2004**, *47*, 6117–6119.

[37] Hanessian, S.; Yang, G.; Rondeau, J. M.; Neumann, U.; Betschart, C.; Tintelnot-Blomley, M. *J. Med. Chem.* **2006**, *49*, 4544–4567.

[38] Hanessian, S.; Yun, H.; Hou, Y.; Yang, G.; Bayrakdarian, M.; Therrien, E.; Moitessier, N.; Roggo, S.; Veenstra, S.; Tintelnot-Blomley, M.; Rondeau, J. M.; Ostermeier, C.; Strauss, A.; Ramage, P.; Paganetti, P.; Neumann, U.; Betschart, C. *J. Med. Chem.* **2005**, *48*, 5175–5190.

Table 2. Summary of Statistics of the Linear Regression Fit

set	I[a]	A[a]	B[a]	size of set	outliers	R^2[b]	Fisher[b]value	σ[b]	Q^2[c]	σ_{cv}[c]	ϕ[d]	κ[d]	$\lambda \times 10^{12}$ (m)[d]
Full	295.407	−16.270	−3.712	66	0	0.51	33.29	2.60	0.46	2.68	2.09	−3.71	6.30
1	824.647	−60.588	−10.455	12	0	0.93	56.78	0.65	0.85	0.81	2.41	−10.45	9.18
2a	270.655	−2.589	−3.377	9	1	0.98	194.43	0.29	0.95	0.44	0.88	−3.38	2.18
2b	723.108	−26.160	−9.178	13	6	0.93	70.92	0.59	0.88	0.68	1.69	−9.18	6.63
3[e]	319.651	0.052	−4.106	20	3	0.84	44.76	1.25	0.77	1.39	–	−4.11	–
3[e]	319.516	0.000	−4.104	20	3	0.84	94.78	1.22	0.80	1.28	–	−4.10	–

[a] See eq 17. [b] Statistics of linear regression fit. [c] leave-one-out cross validation R^2 and σ. [d] See eq 18. [e] Value of A is too low for ϕ and λ to be determined accurately.

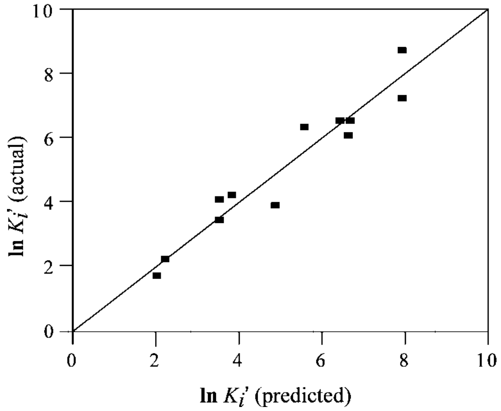

Figure 2. Linear regression fit for bovine trypsin (data set 1).

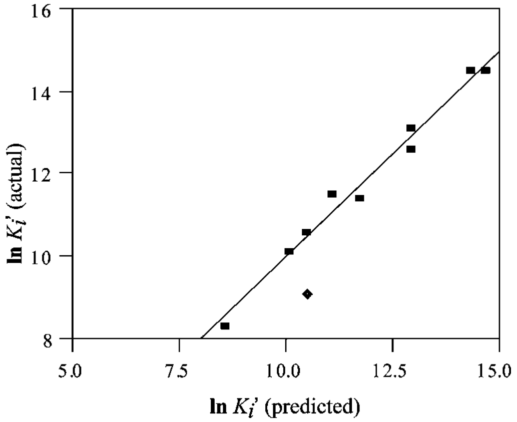

Figure 3. Linear regression fit for the β-secretase data set: 2a with complexes whose assays contained 10% DMSO (◆ - outlier, 2oht).

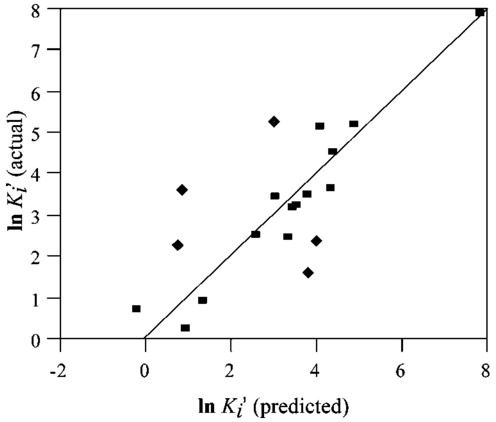

Figure 4. Linear regression fit for the β-secretase data set: 2b with complexes whose assays contained no DMSO (◆ - outliers).

2iqg, 2iso, and 2irz (and 2ph6), use as a substrate a fusion protein containing maltose binding protein at the amino terminal end and 125 amino acids of the amyloid precursor protein amino terminal end. All the other assays use significantly smaller peptide substrates in the assay. The use of a constant χ in the Wilke-Chang equation for the empirical fit could lead to errors in the estimation of D, since its value suggested value varies from 1.61 to 2.9 depending on the chemical nature of the molecule.[25,27] This is a flaw in our analysis since the specific chemical characteristics of the molecule were not considered in the empirical determination of D. The ligands of 1ym1 and 1ym2, for instance, have very little aromatic character to them while that is not true of the other ligands. Also, conditions specific to those assays may also be responsible for the observed deviation. Even so, the deviation in the predicted ln[K_i']s for these outliers is within a factor of 2.5, which implies that the predicted activity is correct to an order of magnitude. In general, since the ligands in this study are significantly more complex than those used to drive the empirical fit in the Wilke-Chang equation, extrapolation to more complex molecules may increase the error.

Aldose Reductase. The aldose reductase data set[40−53] is expected to be most noisy, since the data spans over a decade

(39) Rajapakse, H. A.; Nantermet, P. G.; Selnick, H. G.; Munshi, S.; McGaughey, G. B.; Lindsley, S. R.; Young, M. B.; Lai, M. T.; Espeseth, A. S.; Shi, X. P.; Colussi, D.; Pietrak, B.; Crouthamel, M. C.; Tugusheva, K.; Huang, Q.; Xu, M.; Simon, A. J.; Kuo, L.; Hazuda, D. J.; Graham, S.; Vacca, J. P. *J. Med. Chem.* **2006**, *49*, 7270–7273.

(40) El-Kabbani, O.; Darmanin, C.; Oka, M.; Schulze-Briese, C.; Tomizaki, T.; Hazemann, I.; Mitschler, A.; Podjarny, A. *J. Med. Chem.* **2004**, *47*, 4530–4537.

(41) Wilson, D. K.; Tarle, I.; Petrash, J. M.; Quiocho, F. A. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 9847–9851.

(42) Urzhumtsev, A.; TeteFavier, F.; Mitschler, A.; Barbanton, J.; Barth, P.; Urzhumtseva, L.; Biellmann, J. F.; Podjarny, A. D.; Moras, D. *Structure* **1997**, *5*, 601–612.

(43) Calderone, V.; Chevrier, B.; Van Zandt, M.; Lamour, V.; Howard, E.; Poterszman, A.; Barth, P.; Mitschler, A.; Lu, J. H.; Dvornik, D. M.; Klebe, G.; Kraemer, O.; Moorman, A. R.; Moras, D.; Podjarny, A. *Acta Crystallogr. D* **2000**, *56*, 536–540.

(44) Kinoshita, T.; Miyake, H.; Fujii, T.; Takakura, S.; Goto, T. *Acta Crystallogr. D* **2002**, *58*, 622–626.

(45) Howard, E. I.; Sanishvili, R.; Cachau, R. E.; Mitschler, A.; Chevrier, B.; Barth, P.; Lamour, V.; Van Zandt, M.; Sibley, E.; Bon, C.; Moras, D.; Schneider, T. R.; Joachimiak, A.; Podjarny, A. *Proteins-Struct. Funct. Bioinf.* **2004**, *55*, 792–804.

(46) Ruiz, F.; Hazemann, I.; Mitschler, A.; Joachimiak, A.; Schneider, T.; Karplus, M.; Podjarny, A. *Acta Crystallogr. D* **2004**, *60*, 1347–1354.

(47) Van Zandt, M. C.; Jones, M. L.; Gunn, D. E.; Geraci, L. S.; Jones, J. H.; Sawicki, D. R.; Sredy, J.; Jacot, J. L.; DiCioccio, A. T.; Petrova, T.; Mitschler, A.; Podjarny, A. D. *J. Med. Chem.* **2005**, *48*, 3141–3152.

(48) El-Kabbani, O.; Darmanin, C.; Schneider, T. R.; Hazemann, I.; Ruiz, F.; Oka, M.; Joachimiak, A.; Schulze-Briese, C.; Tomizaki, T.; Mitschler, A.; Podjarny, A. *Proteins-Struct. Funct. Bioinf.* **2004**, *55*, 805–813.

(49) Steuber, H.; Zentgraf, M.; Podjarny, A.; Heine, A.; Klebe, G. *J. Mol. Biol.* **2006**, *356*, 45–56.

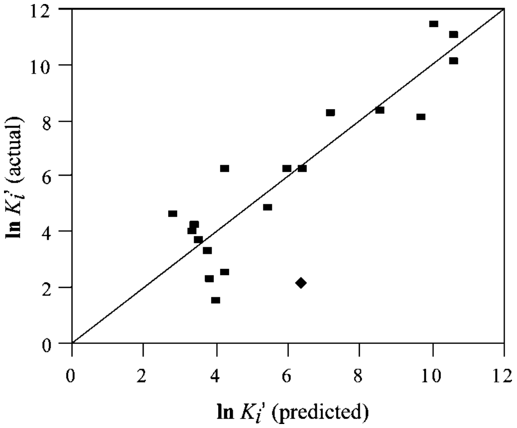

Figure 5. Linear regression fit for the aldose reductase data set 3. (◆ - outlier, only 1pwl is shown.)

of work and is obtained from the work of several different laboratories. In fact an examination of the assays shows that the affinity data was obtained from the use of a wide range of substrates (glucose, xylitol, xylose, glyceraldehydes, and benzyl alcohol). The LHS of eq 18 involves [E] whose value depends on [E_{Total}] and the ratio of [S]/K_m, where [S] is the substrate concentration. Assays using different substrates using varying ratios of [S]/K_m are therefore likely to introduce further noise into the data set beyond that due to experimental error. In fact, for two of the inhibitors, different values of affinity separated by an order of magnitude have been reported by different groups; 44[44] and 400[54] nM for zenarestat (1iei) and, 3[54] and 60[54] nM for zopolrestat (1mar). We used the more recent estimates in our analysis for consistency.

Since eq 18 predicts the affinities to very high precision, noisy data is expected to lead to inaccurate empirical prediction formulas for the activity. This is observed in the fit for this data set (figure 5), where the contribution of the TermA is not captured at all. This data set has three outliers: 1pwl, 1z8a, 1z89. While we can not explain the inaccuracy in the prediction of 1pwl, an examination of the assay buffers of the ligands of 1z89 and 1z8a, not surprisingly, found 20% DMSO in the assay buffer. Here, ($\sigma = 1.22$) activities are predicted to within an order of magnitude, which is close to the expected level of noise in this data set. The widely varying chemical composition of the ligands in this data set will also increase errors in the evaluation of the diffusivities. A high sensitivity to the diffusivity is expected in the prediction of the activities due to the nature of eq 18, with the sensitivity being much higher for TermA than TermB, since a D^2 features in TermA while TermB involves ln ($1/D^3$).

Discussion

In general it appears that eq 18 can predict free energies almost to the precision of the available data for a linear regression fit. Of great interest is the effect of the magnitude of the diffusion coefficient on the observed inhibitor affinities. The systematic error observed in the residuals in Figure 1 (Table 1) appears to be due to the different correction factor (ϕ) that was required for different enzyme assays. The correction factor, ϕ, was determined by equalizing the coefficients of TermA and TermB in the linear regression fit (Table 2). The value of ϕ explains the strong dependence on TermB in data set 2a. This appears to be due to the lower diffusivities of these molecules in comparison to the trypsin data set. Larger diffusivities of the ligands in the trypsin data set reduce the accuracy of prediction with TermB alone in the trypsin data set ($R^2 = 0.39$, $\sigma = 1.55$, F=5.84). The decrease in the diffusivities of data set 2a is most likely due to the increased viscosity of the solution due to addition of DMSO, since viscosity appears in the denominator of Wilke-Chang equation. The sensitivity of the activity of the molecules to their diffusion coefficients highlights the importance of the conditions of the assay toward the observed activity.

An interesting implication of the strong dependence on TermB for subset 2a is that the energetics of the binding are of little consequence for predicting activity for such systems. The activity almost entirely depends on the diffusion coefficient. The Wilke-Chang equation implies that activity will be strongly correlated with molecular volume. Since molecular masses and volumes are strongly correlated ($R^2 = 0.96$ for the entire data set), the molecular mass would also be highly correlated with activity. This interesting fact has been observed by Kim and Skolnick,[55] where a comparison of scoring functions of different docking algorithms led to the observation that molecular mass was as good a predictor of activity as the scoring functions themselves. There, the authors argue that this observation may be due to the fact that most rational drug design involves improvement of efficacy of drug leads through the addition of favorable functional moeties, thus leading to an increase in mass. Equation 18 implies that the reason for the observed correlation is more fundamental since the size of the molecule directly affects its diffusivity. In fact we find that the use of the logarithm of the molecular weight (ln(M_W)) and k/M_W (where k is the trace of the Hessian matrix) resulted in empirical correlations with similar accuracies to those achieved by eq 18.

Apart from the composition of the assay buffer, other factors such as temperature also strongly influence D. For instance, from the Wilke-Chang equation, the ratio of diffusivities of a molecule at infinite dilution at temperatures 310 and 298 would be 1.34. Thus, for a low molecular weight ligand, eq 18 would predict significantly different activities at different temperatures.

One common assumption made to simplify the configurational integral of eq 4 is that it can be approximated by “end-point” calculations that limit the computation to the solvated bound/unbound ligand/receptor alone. Equation 18 implies that such an approximation will capture only a part of the picture, neglecting the contribution of the molecules’ diffusivity toward binding affinity.

Despite the success of eq 18 with the trypsin and β-secretase data sets and the fairly low computational cost of its evaluation, several problems exist with its usage on a large scale for predicting activities. The first comes from the limitations of the empirical nature of the Wilke-Chang equation for predicting diffusivities. It is not surprising therefore, that the highest accuracy in the prediction of activities was achieved for data sets 1 and 2a where the ligands were of similar chemical nature. For a data set with molecules of widely differing properties,

[50] Brownlee, J. M.; Carlson, E.; Milne, A. C.; Pape, E.; Harrison, D. H. T. *Bioorg. Chem.* **2006**, *34*, 424–444.

[51] Steuber, H.; Zentgraf, M.; Gerlach, C.; Sotriffer, C. A.; Heine, A.; Klebe, G. *J. Mol. Biol.* **2006**, *363*, 174–187.

[52] Petrova, T.; Ginell, S.; Mitschler, A.; Hazemann, I.; Schneider, T.; Cousido, A.; Lunin, V. Y.; Joachimiak, A.; Podjarny, A. *Acta Crystallogr. D* **2006**, *62*, 1535–1544.

[53] Steuber, H.; Heine, A.; Klebe, G. *J. Mol. Biol.* **2007**, *368*, 618–638.

[54] Ehrig, T.; Bohren, K. M.; Prendergast, F. G.; Gabbay, K. H. *Biochemistry* **1994**, *33*, 7157–7165.

[55] Kim, R.; Skolnick, J. *J. Comput. Chem.* **2008**, *29*, 1316–1331.

such as data set 3, the use of a single empirical coefficient will most likely lead to significant errors in the prediction of diffusion coefficients. Besides, the success of the Wilke-Chang equation has been demonstrated for a limited number of small organic molecules in solution and it is not clear if its extrapolation to larger and chemically more diverse molecules typical of drug candidates would lead to accurate predictions of diffusion coefficients. It appears therefore that independent methods for determining diffusivities may have to supplement drug design experiments.

A second difficulty of the method outlined here is with the determination of the trace of the Hessian matrix. The trace is very sensitive to the final minimized complex as well as to the minimization protocol used. The correlation coefficient, R^2, of the linear regression fit for trypsin data set decreases significantly when cut-offs (8 Å) were used for evaluating the nonbonded interactions. A Generalized-Born solvation model was necessary during minimization to simulate the effect of the solvent. This is expected since the minimum of V_{eff} is for the solvated complex. Explicit solvation could likely improve the fit even more, but the computational costs would quickly become prohibitive. Careful attention must be paid to the protonation states of the ligands and any titrable protein side-chains. For instance, 1o36 and 1o3g fit only after protonation of a hydroxyl group.

Also interesting is the relationship that results for V_{eff}, where the significant factor for unbinding is not the interaction potential as intuition suggests, but its second derivative. In other words, the unbinding free energy depends not on the depth of the potential well that the ligand finds itself in, but rather on the curvature of this potential in the vicinity of energy minimum. This is a consequence of the overdamping effect of the solvent that appears as an assumption in the Fokker-Plank equation.

Our greatest challenge in this work, was to find data sets of sufficiently large size (for statistical significance) that met all of our criteria, namely the availability of affinity data together with crystal structures for every complex, as well as the absence of significant conformational rearrangements in the active-site associated with binding. The absence of metal atoms and significant interactions with crystal waters was also considered essential as we were interested in the simplest possible data sets to test the theory. We believe that it is due to this choice that we obtained reasonably good fits despite significant simplification of the physics. It is quite likely that other rate-limiting steps such as secondary binding events prior to unbinding or conformational rearrangements in the active-site during unbinding would significantly complicate the picture and perhaps demand a less trivial solution.

Two other methods for affinity prediction that were derived involving assumptions, albeit very different from ours, that lead to a simplification of the underlying physics and hence a reduction of the computational expense can be mentioned here. The first of these is the Linear Interaction Energy (LIE) method developed by Aqvist and co-workers.[56] The LIE method was derived as an approximation to free energy perturbation, where a linear response is assumed for electrostatic interactions, together with an empirical expression for nonpolar effects. The interaction energies are evaluated from the average of molecular dynamics simulation (MD) energies. The second method is the Mining Minima method developed by Gilson and co-workers[57] and applies to the direct computation of the configurational integral of a molecule (eq 4) as a sum of the contribution of low energy states using a Monte Carlo (MC) technique. Both methods require conformational sampling through MD or MC techniques. In comparison, our method has the advantage that a single conformation is sufficient for free energy prediction.

Conclusions

In a recent review of docking and scoring functions, Leach et al.[58] point out that the accuracy of scoring functions has reached a plateau and that there is need for a breakthrough to develop. The path-integral formalism presented in this work offers an alternate perspective on the problem of binding free energy prediction.

The practical complexity of the traditional statistical thermodynamics approach implies that the effect of the assay buffer cannot be reasonably accounted for without significantly increasing the already intractably high computational time of estimating the configurational integral of eq 4. The path integral formalism presented here is elegant not only for its high accuracy, but also for its computational expense, which is minimal.

In its current form, eq 18 has great potential for the development of improved scoring functions for docking algorithms. A more extensive verification of eq 18 is certainly necessary, and at this point this is mainly limited by lack of high quality data sets such as that of bovine trypsin. It remains to be seen how eq 18 performs in real life drug design scenarios, where typically only a single bound complex is available as a template for the prediction of other bound complexes through docking. It would also greatly help if inhibition experiments were supplemented with experimental determination of the ligands' diffusivities in their respective buffers, which would remove the reliance on empirical relations such as Wilke-Chang equation that was used in this work.

Acknowledgment. We thank Dr. Abdallah Sayyed-Ahmad and Dr. Anthony D. Hill for helpful discussions. This work was supported by a grant from NIH (GM 070989). Computational support from the Minnesota Supercomputing Institute (MSI) is gratefully acknowledged. This work was also partially supported by National Computational Science Alliance under MCA04T033.

Supporting Information Available: Contains derivations of important relations in the kinetics of competitive inhibition. This material is available free of charge via the Internet at http://pubs.acs.org.

JA807460S

[56] Hansson, T.; Marelius, J.; Åqvist, J. *J. Comput. Aided Mol. Des.* **1998**, *12*, 27–35.

[57] Gilson, M. K.; Given, J. A.; Head, M. S. *Chem. Biol.* **1997**, *4*, 87–92.

[58] Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. *J. Med. Chem.* **2006**, *49*, 5851–5855.