

Paper **COMP 160** presented at the
228th National Meeting of the American Chemical Society,
August 22–26, 2004, Philadelphia, PA

Method for Computing Protein Binding Affinity^{*†}

Charles F. F. Karney[‡] and Jason E. Ferrara
Sarnoff Corporation, Princeton, NJ 08543-5300

Stephan Brunner[§]
Locus Pharmaceuticals, Inc., Blue Bell, PA 19422-2700

Abstract

A Monte Carlo method is given to compute the binding affinity of a ligand to a protein. The method involves extending configuration space by a discrete variable indicating whether the ligand is bound to the protein and a special Monte Carlo move which allows transitions between the unbound and bound states. Provided that an accurate protein structure is given, that the protein-ligand binding site is known, and that an accurate chemical force field together with a continuum solvation model is used, this method provides a quantitative estimate of the free energy of binding.

*This work was supported, in part, by the U.S. Army Medical Research and Materiel Command under Contract No. DAMD17-03-C-0082.

†A detailed description of this work is available as a preprint at:

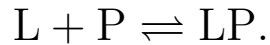
<http://arxiv.org/cond-mat/0401348>

‡E-mail: ckarney@sarnoff.com

§Present address: CRPP, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland.

Introduction

Consider a drug molecule, the ligand L, binding reversibly to a protein P, via



In equilibrium, reaction is governed by the dissociation constant

$$K_d = \frac{[L][P]}{[LP]}.$$

Define the binding affinity as

$$pK_d = -\log_{10} \left(\frac{K_d/N_A}{1 \text{ kmol m}^{-3}} \right),$$

where N_A is the Avogadro constant.

Goal of this work: given

- protein structure
- ligand binding site
- chemical force field
- solvation model

compute K_d .

Benefits are

- screen drug leads prior to synthesis
- guide the design of drug leads

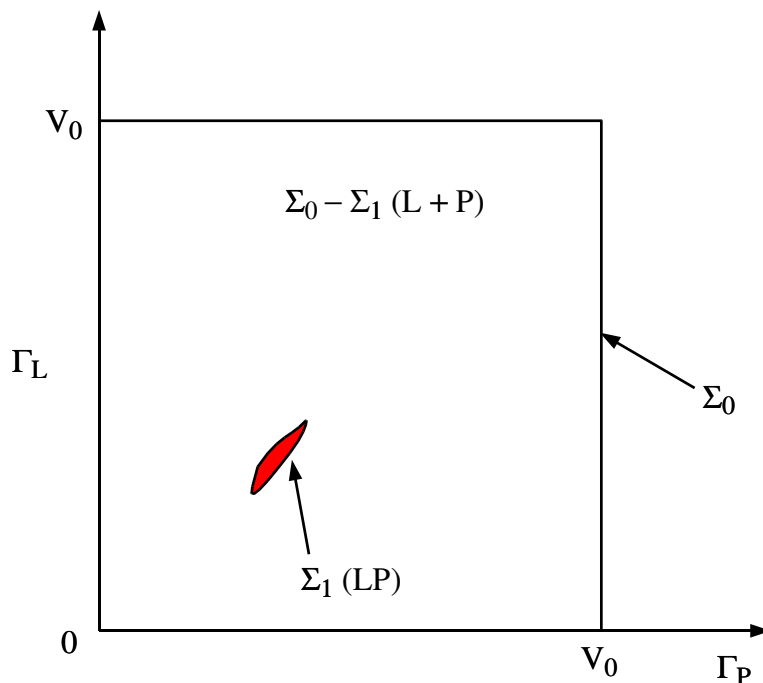


Figure 1: Configuration space showing the volumes corresponding to the unbound molecules $L + P$ and the complex LP .

Formulation

Consider a system of volume V_0 consisting of a ligand molecule L and a protein molecule P in a solvent. The state of the system is given by $\Gamma = [\Gamma_L, \Gamma_P]$, where Γ_M represents the phase space configuration of molecule M (position, orientation, conformation); see Fig. 1. In equilibrium, the system obeys the Boltzmann distribution¹

$$f(\Gamma) \propto \exp[-\beta E(\Gamma)],$$

where $\beta = 1/(kT)$, T is the temperature, and k is the Boltzmann constant, and $E(\Gamma)$ is the energy of the system.

Σ_0 corresponds to all allowable configurations; Σ_1 corresponds to the LP complex. In the dilute limit $V_0 \rightarrow \infty$, we find^{2,3,4}

$$K_d = \frac{1}{V_0} \frac{\int_{\Sigma_0} \exp[-\beta E_0(\Gamma)] d\Gamma}{\int_{\Sigma_1} \exp[-\beta E_1(\Gamma)] d\Gamma}, \quad (1)$$

where $E_1(\Gamma)$ is the full energy of the system and $E_0(\Gamma)$ is the “unbound” energy, ignoring the interaction between L and P.

Combine the bound and unbound systems by extending phase space

$$\Gamma \rightarrow [\Gamma, \lambda]; \quad \lambda = \{0, 1\}.$$

Define canonical average by

$$\langle X \rangle = \frac{\sum_{\lambda} \int d\Gamma \exp[-\beta E_{\lambda}(\Gamma)] X_{\lambda}(\Gamma)}{\sum_{\lambda} \int d\Gamma \exp[-\beta E_{\lambda}(\Gamma)]}.$$

Now Eq. (1) can be rewritten as

$$K_d = \frac{1}{V_0} \frac{\langle \delta_{\lambda 0} \rangle}{\langle \delta_{\lambda 1} \rangle}, \quad (2)$$

where $\delta_{\lambda\mu}$ is the Kronecker delta. Because the definition of K_d is independent of V_0 , we can pick V_0 to give $\langle \delta_{\lambda 0} \rangle \sim \langle \delta_{\lambda 1} \rangle$.

Wormhole Monte Carlo

We can compute the canonical averages in Eq. (2) using the Monte Carlo method⁵ to make steps from $[\Gamma, \lambda]$ to $[\Gamma', \lambda']$ with acceptance probability

$$\min[1, \exp(-\beta[E_{\lambda'}(\Gamma') - E_{\lambda}(\Gamma)])].$$

However, the estimate of the K_d will be very poor, because transitions between $\lambda = 0$ and 1 will be extremely rare; see Fig. 2.

Remedy this problem by restricting the standard moves to changes in Γ only, and allowing changes in λ via “wormhole moves” which connect otherwise disconnected regions of configuration space.

Wormhole moves combine two concepts

- squeezing Γ space is equivalent to decreasing E (via Jacobian factor when transforming integrals),^{6,7}
- allowing direct jumps between wells (subject to “detailed balance”).^{8,9}

Define a set of wormhole regions w, w', w'', \dots . These are subsets of $[\Gamma, \lambda]$ space with configuration space volumes of v, v', v'', \dots . Intersperse wormhole moves with regular Monte Carlo moves.

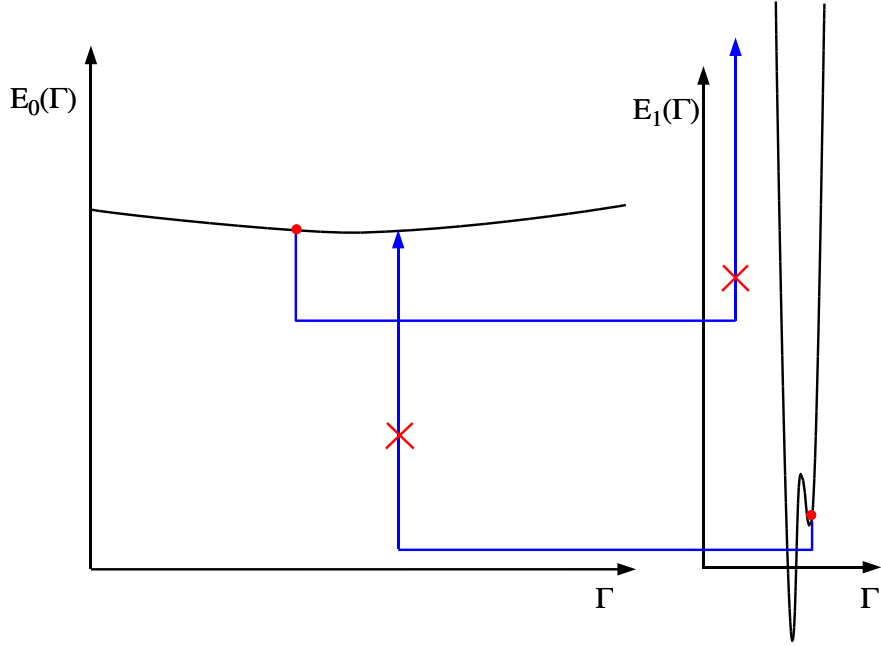


Figure 2: Schematic representation of the $E_0(\Gamma)$ (shallow and wide) and $E_1(\Gamma)$ (deep and narrow). Conventional Monte Carlo moves between $\lambda = 0$ and 1 (shown as blue lines) are nearly always rejected because they lead to a large increase in energy.

Define a wormhole move as follows:

- pick random wormhole w ;
- if $[\Gamma, \lambda] \notin w$, **reject the move**;
- pick random wormhole w' ;
- pick $[\Gamma', \lambda']$ uniformly in w' ;
- compute $\Delta E = E_{\lambda'}(\Gamma') - E_{\lambda}(\Gamma)$;
- with prob. $\exp(-\beta\Delta E)v'/v$, **accept the move**:

$$[\Gamma_{\text{new}}, \lambda_{\text{new}}] = [\Gamma', \lambda'];$$

- otherwise, **reject move**:

$$[\Gamma_{\text{new}}, \lambda_{\text{new}}] = [\Gamma, \lambda].$$

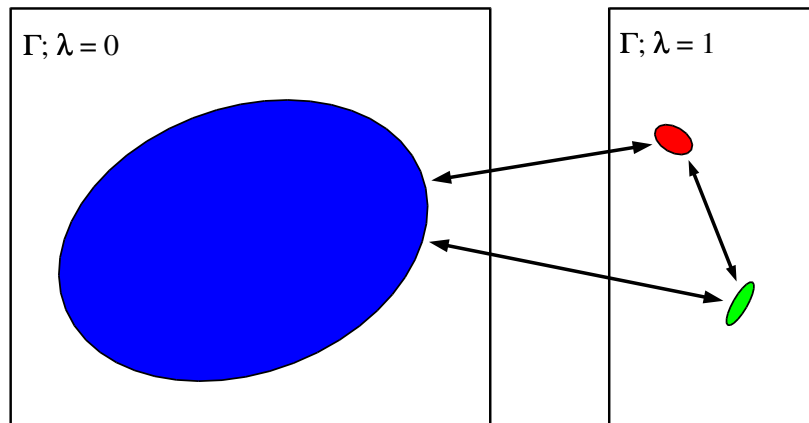


Figure 3: Typical wormholes for the case illustrated in Fig. 2. The large ratio of the volume of the unbound wormhole compared to the bound wormholes compensates for the higher energy of the unbound configurations. This results in accepted wormhole moves between all the wormholes.

If the mean energy of configurations in w scales as $\beta^{-1} \ln v + \text{const.}$,

the acceptance probability is $O(1)$;

see Fig. 3.

Finding the wormhole

In order for the wormhole method to be practical, we need a reliable way of choosing the wormholes. For simplicity take protein to be rigid and fixed. Fix the bond lengths and bond angles in the ligand and allow l bonds to rotate. Configuration of system is given by

- position of L (3),
- orientation of L (3),
- conformation of L (l).
- Total dimensionality is $n = l + 6$.

Carry out canonical Monte Carlo simulations with $E_\lambda(\Gamma)$ separately for $\lambda = 0$ and 1. For each λ , obtain a canonical set of configurations $\{\Gamma\}$. Fit n -dimensional ellipsoid to each $\{\Gamma\}$ with center at the mean configuration $\langle\Gamma\rangle$. Compute the deviation from the mean, $\delta\Gamma = \Gamma - \langle\Gamma\rangle$, and compute a covariance matrix

$$\langle\delta\Gamma \delta\Gamma\rangle = \mathbf{B}\mathbf{B}^T.$$

We take the semi-axes of the ellipsoid to be the columns of $\sqrt{n}\mathbf{B}$.

Ellipsoids are a natural choice to use to fit the set of configurations:

- The iso-density contours of the distribution in a harmonic well are ellipsoids.
- It is easy to sample points randomly from an ellipsoid.
- Conversely, it is easy to test that a point lies inside an ellipsoid.
- The volume of an n -dimensional ellipsoid is given by

$$v_n = \frac{\pi^{n/2}}{(n/2)!} \prod_{i=1}^n a_i,$$

where a_i is the length of i th semi-axis.

Test suitability of ellipsoid by demanding that $O(1)$ of the configurations sampled uniformly from it have energies close to its mean energy $\langle E_\lambda(\Gamma) \rangle$. If test fails, split $\{\Gamma\}$ into two sets according to the sign of $\delta\Gamma$ projected along the largest semi-axis of the ellipsoid and construct new trial wormholes from each of these sets.

Method of finding wormholes depends on the samples “spanning” a volume of phase space. Requires that the dimensionality of phase space be sufficiently small. Thus

- **Must use an implicit solvation model.**
- **Minimize number of degrees of freedom for conformational changes by fixing the bond lengths and bond angles.**

How good are wormhole moves?

Example: *p*-amino-benzamidine (Fig. 4) bound to trypsin:

- Protein structure from trypsin-benzamidine complex, 1BTY¹⁰ (Fig. 5).
- At physiological pH, ligand is protonated (net charge of +1).
- Amber 7 force field^{11,12,13} and GB/SA solvation model.^{14,15,16}
- Find 16 unbound and 8 bound wormholes.
- Pick $V_0 = 0.39 \times 10^{-18} \text{ m}^3$.
- Binding affinity calculation of 5×10^6 steps.

Data from binding affinity calculations. For every 1000 steps:

| | |
|--|-----|
| wormhole attempts | 900 |
| $\Gamma \in w?$ | 26 |
| successful wormhole move | 15 |
| λ transition $0 \rightarrow 1 \rightarrow 0$ | 3 |

Main cost is evaluation of $E_1(\Gamma)$. But the frequent $\Gamma \notin w$ steps are free! Thus

$$\text{Cost of each } \lambda \text{ transition} \sim 3 \times E_1(\Gamma)$$

Wormholes are *effective* in making the transition between the bound and unbound systems.

Computational result: $\text{p}K_d = 7.99 \pm 0.01$

Experimental^{17,18} result: $\text{p}K_d = 5.1 \pm 0.1$

Discrepancy may be accounted for by modest, $\sim 20\%$, errors in the force field. Protein flexibility probably also be important.

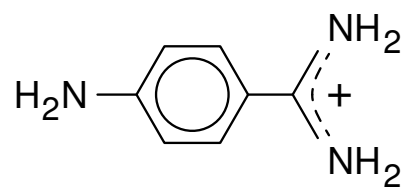


Figure 4: The structure of *p*-amino-benzamidine.

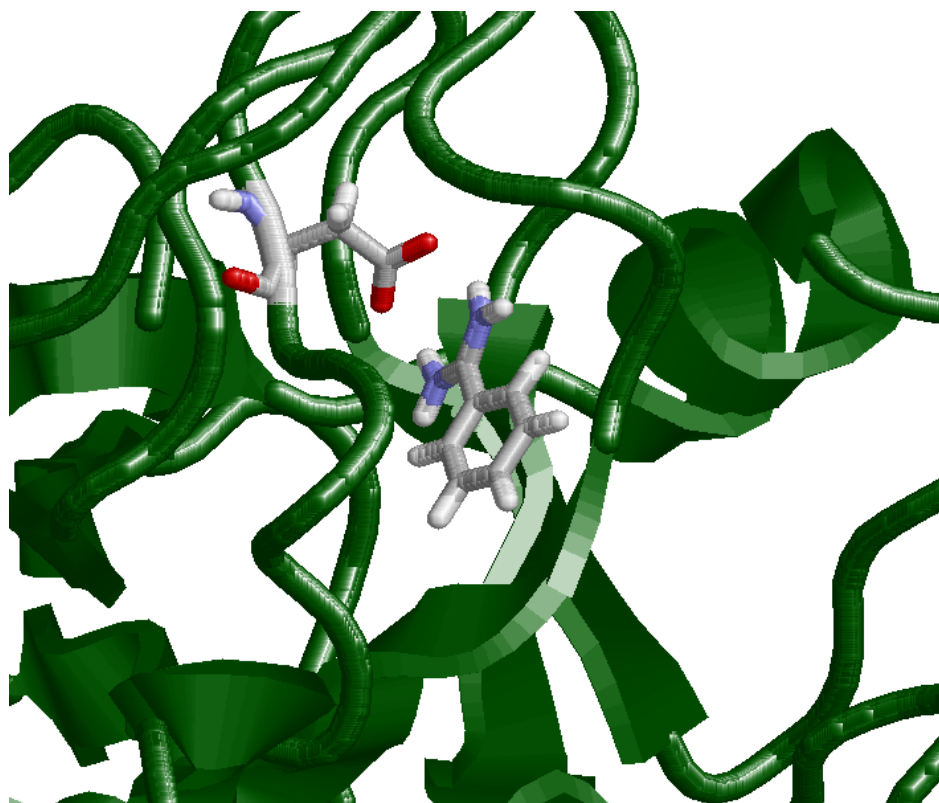


Figure 5: Trypsin-benzamidine complex, 1BTY.

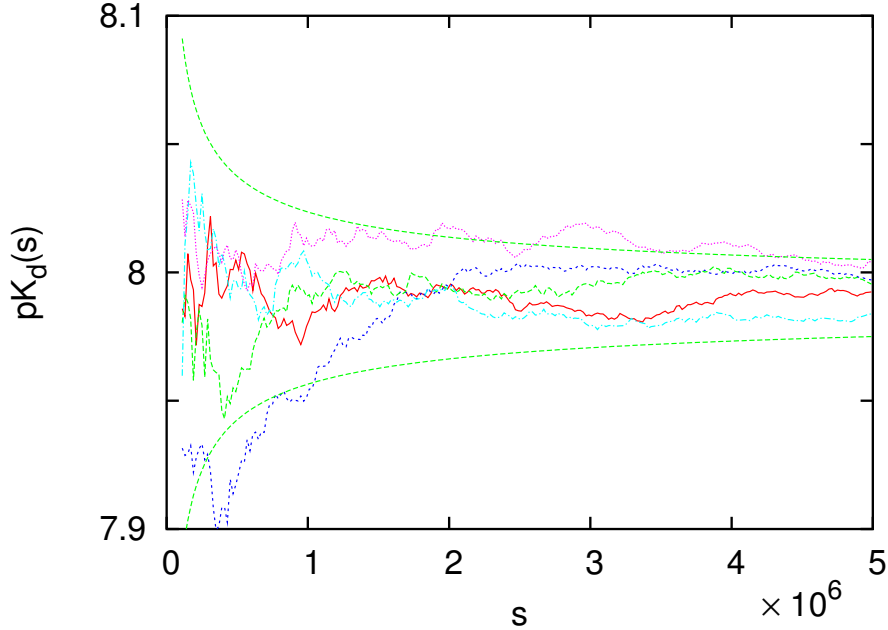


Figure 6: Cumulative estimates $pK_d(s)$ obtained from the first s steps of 5 independent Monte Carlo runs. The dashed lines show convergence as $1/\sqrt{s}$ to the mean value of 7.99.

Convergence

Repeat computation 5 times and plot the cumulative estimates of pK_d based on the first s steps of each run; see Fig. 6. The convergence depends on how rapidly the switch between $\lambda = 0$ and 1 is made. This is determined by the λ -correlation function

$$C_t = \langle (\lambda_s - p)(\lambda_{s+t} - p) \rangle_s,$$

where $p = \langle \delta_{\lambda 1} \rangle$, λ_s is the value of λ at simulation step s and $\langle \dots \rangle_s$ denotes an average over steps. Figure 7 shows C_t for several different values of V_0 . In a Markov chain of length s , the expected number

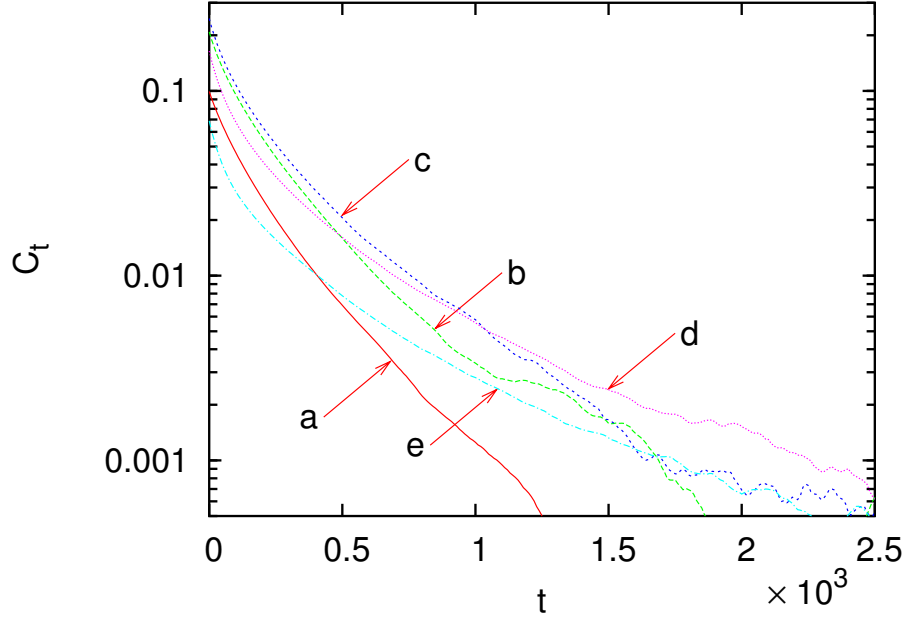


Figure 7: The λ -correlation function C_t . The curves show C_t for $V_0/(0.39 \times 10^{-18} \text{ m}^3) =$ (a) $\frac{1}{10}$, (b) $\frac{1}{3}$, (c) 1, (d) 3, and (e) 10.

of bound states is ps , while, for $s \rightarrow \infty$, the variance in the number of bound states is $2Ds$, where

$$D = \frac{1}{2}C_0 + \sum_{t>0} C_t = \frac{1}{2}C_0\tau.$$

This provides us with the definition of the correlation time, τ . From Fig. 7, we see that C_t decays approximately exponentially so that the sum converges giving $\tau \sim 350$.

Problem of determining pK_d is equivalent to determining the bias of a coin where probability of heads is p ; see Fig. 8.

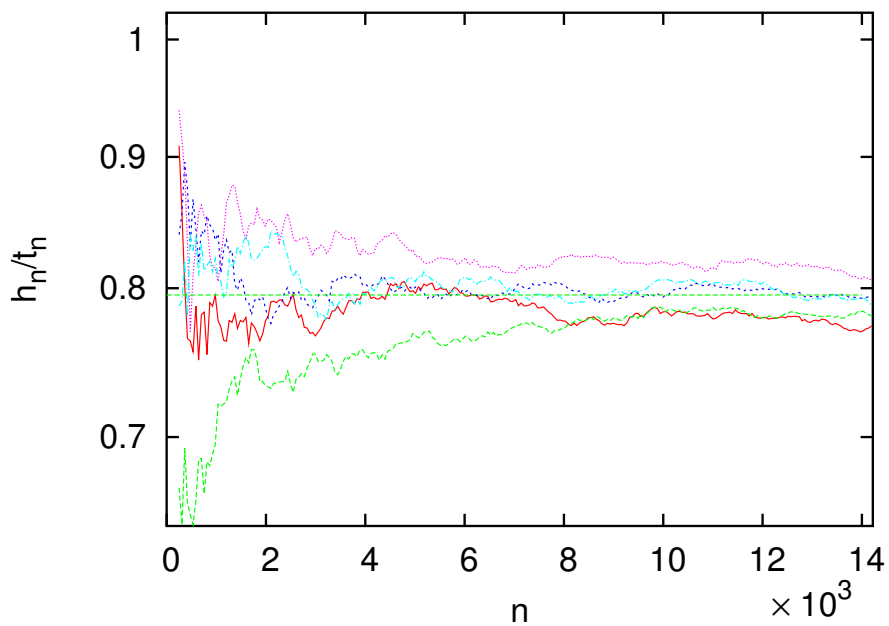


Figure 8: Estimation of the bias of a coin with $p = 0.443$. Five independent runs are shown; h_n (t_n) is the number of heads (tails) observed in n trials. The axes have been adjusted to allow the figure to be directly compared with Fig. 6.

Discussion

- Direct sampling of two physical systems, $E_1(\Gamma)$ and $E_0(\Gamma)$ (compare with free energy perturbation theory).
- Accumulating 1 (in either $\langle \delta_{\lambda 0} \rangle$ or $\langle \delta_{\lambda 1} \rangle$)
 - no rare large values (high error)
 - no frequent small values (high cost)
- Can sample with $E_\lambda^*(\Gamma) \approx E_\lambda(\Gamma)$ and compensate in the canonical averages.
- Can apply standard Monte Carlo techniques
 - preferential sampling
 - early rejection
 - force bias
 - etc.
- Can extend to treat
 - limited protein flexibility
 - ring flexibility in ligand
 - protonation states (at constant pH)
 - tautomers
- **Limitations**
 - explicit solvent not treated
 - only as good as crystal structure and force field

References

- [1] L. D. Landau and E. M. Lifshitz, *Statistical Physics*; Vol. 5 of *Course of Theoretical Physics*; Pergamon Press, 2nd ed., 1969.
- [2] M. Mezei and D. L. Beveridge, *Ann. N.Y. Acad. Sci.*, **1986**, 482, 1–23.
- [3] C. H. Bennett, *J. Comp. Phys.*, **1976**, 22, 245–268.
- [4] H. Luo and K. Sharp, *Proc. Nat. Acad. Sci.*, **2002**, 99, 10399–10404.
- [5] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *J. Chem. Phys.*, **1953**, 21, 1087–1092.
- [6] M. A. Miller and W. P. Reinhardt, *J. Chem. Phys.*, **2000**, 113, 7035–7046.
- [7] Z. Zhu, M. E. Tuckerman, S. O. Samuelson, and G. J. Martyna, *Phys. Rev. Lett.*, **2002**, 88, 100201.
- [8] A. F. Voter, *J. Chem. Phys.*, **1985**, 82, 1890–1899.
- [9] H. Senderowitz, F. Guarnieri, and W. C. Still, *J. Am. Chem. Soc.*, **1995**, 117, 8211–8219.
- [10] B. A. Katz, J. Finer-Moore, R. Mortezaei, D. H. Rich, and R. M. Stroud, *Biochemistry*, **1995**, 34, 8264–8280; URL <http://www.rcsb.org/pdb/cgi/explore.cgi?pdbId=1bty>.
- [11] D. A. Case, D. A. Pearlman, J. W. Caldwell, T. E. Cheatham, III, J. Wang, W. S. Ross, C. L. Simmerling, T. A. Darden, K. M. Merz, R. V. Stanton, *et al.*; *Amber 7*; University of California, San Francisco, **2002**; URL <http://amber.scripps.edu/doc7/>.
- [12] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, *J. Am. Chem. Soc.*, **1995**, 117, 5179–5197.
- [13] C. I. Bayley, P. Cieplak, W. D. Cornell, and P. A. Kollman, *J. Phys. Chem.*, **1993**, 97, 10269–10280.
- [14] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, *J. Am. Chem. Soc.*, **1990**, 112, 6127–6129.
- [15] G. D. Hawkins, C. J. Cramer, and D. G. Truhlar, *Chem. Phys. Lett*, **1995**, 246, 122–129.
- [16] V. Tsui and D. A. Case, *J. Am. Chem. Soc.*, **2000**, 122, 2489–2498.
- [17] M. Mares-Guia and E. Shaw, *J. Biol. Chem.*, **1965**, 240, 1579–1585.
- [18] S. M. Schwarzl, T. B. Tschopp, J. C. Smith, and S. Fischer, *J. Comp. Chem*, **2002**, 23, 1143–1149.