Biophysical Journal, Volume 98

**Supporting Material**

# Accounting for ligand conformational restriction in calculations of protein-ligand binding affinities

Cen Gao, Min-Sun Park, and Harry A. Stern

# Supplementary Material

## Calculating the harmonic configuration integral

$$Z_h \approx 8\pi^2 V e^{-\beta U(\mathbf{q}_0)} \int e^{-\frac{1}{2}\Delta q^T \mathbf{H} \Delta q} \Big( J_0 + \sum_{i=1}^{3N-6} J_i \Delta q_i + \frac{1}{2} \sum_{i,j=1}^{3N-6} J_{ij} \Delta q_i \Delta q_j + \frac{1}{6} \sum_{i,j,k=1}^{3N-6} J_{ijk} \Delta q_i \Delta q_j \Delta q_k + \frac{1}{24} \sum_{i,j,k,l=1}^{3N-6} J_{ijkl} \Delta q_i \Delta q_j \Delta q_k \Delta q_l \Big) d^{3N-6} \Delta q. \quad (1)$$

If $\mathbf{H}$ is symmetric and positive definite,

$$\int e^{-\frac{1}{2}\Delta \mathbf{q}^T \mathbf{H} \Delta \mathbf{q}} d^{3N-6} \Delta q = \sqrt{\frac{(2\pi)^{3N-6}}{\det \mathbf{H}}} \quad (2)$$

$$\int \Delta q_i e^{-\frac{1}{2}\Delta \mathbf{q}^T \mathbf{H} \Delta \mathbf{q}} d^{3N-6} \Delta q = 0 \quad (3)$$

$$\int \Delta q_i \Delta q_j e^{-\frac{1}{2}\Delta \mathbf{q}^T \mathbf{H} \Delta \mathbf{q}} d^{3N-6} \Delta q = \sqrt{\frac{(2\pi)^{3N-6}}{\det \mathbf{H}}} H^{-1}_{ij} \quad (4)$$

$$\int \Delta q_i \Delta q_j \Delta q_k e^{-\frac{1}{2}\Delta \mathbf{q}^T \mathbf{H} \Delta \mathbf{q}} d^{3N-6} \Delta q = 0 \quad (5)$$

$$\int \Delta q_i \Delta q_j \Delta q_k \Delta q_l e^{-\frac{1}{2}\Delta \mathbf{q}^T \mathbf{H} \Delta \mathbf{q}} d^{3N-6} \Delta q = \sqrt{\frac{(2\pi)^{3N-6}}{\det \mathbf{H}}} \times \left[ H^{-1}_{ij} H^{-1}_{kl} + H^{-1}_{ik} H^{-1}_{jl} + H^{-1}_{il} H^{-1}_{jk} \right]. \quad (6)$$

Therefore, up to fourth order,

$$Z_h \approx 8\pi^2 V e^{-\beta U(\mathbf{q}_0)} \sqrt{\frac{(2\pi)^{3N-6}}{\det \mathbf{H}}} \times \left( J_0 + \frac{1}{2} \sum_{i,j=1}^{3N-6} J_{ij} H^{-1}_{ij} + \frac{1}{24} \sum_{i,j,k,l=1}^{3N-6} J_{ijkl} \left[ H^{-1}_{ij} H^{-1}_{kl} + H^{-1}_{ik} H^{-1}_{jl} + H^{-1}_{il} H^{-1}_{jk} \right] \right). \quad (7)$$

## Details of numerical tests for simple model systems

The potential energy was given by a sum over bond stretches, angle bends and dihedral torsions:

$$U = U_{\text{stretch}} + U_{\text{bend}} + U_{\text{torsion}} = \sum_{\text{stretches}} \frac{1}{2} k_a (\ell - \ell_0)^2 + \sum_{\text{bends}} \frac{1}{2} k_b (\theta - \theta_0)^2 + \sum_{\text{torsions}} \sum_n \frac{1}{2} V_{n,\text{dih}} [1 + \cos n\omega]. \quad (8)$$

Here $k_a$ and $k_b$ are force constants for stretches and bends respectively; $\ell$ and $\theta$ are bond lengths and angles, and $\ell_0$ and $\theta_0$ are their equilibrium values; $\omega$ are dihedrals; and $V_{n,\text{dih}}$ are coefficients for the $n$th term in a Fourier series for each torsion.

The Bennett acceptance ratio method was used to perform the FEP calculation. All molecules were constructed using bond-angle-torsion coordinates as described above. Monte Carlo simulations were performed in internal coordinate space. For water, all internal coordinates (two stretches and a bend) were

used in the covariance matrix calculation. For butane, three bond stretches and two bends were used. The C-C-C-C dihedral torsion was excluded from the QH model since rotation about that dihedral is relatively unhindered and does not contribute a large amount to the total energy. (Note that this torsion *is* included for the analytical calculation.) Six independent runs were performed for each harmonic simulation using the same covariance matrix, with different random seed chosen for the Monte Carlo runs. All calculations were performed at 298.15 K.

## Details of protein-ligand complex test set selection

All entries were first divided into subsets by protein. Duplicate ligands in a subset were removed. Subsets with fewer than five unique ligands were excluded. For a given subset, all complexes were required to have either experimentally determined inhibition constants ($K_i$ values), or dissociation constants ($K_d$ values), following Kim and Skolnick (*J. Comp. Chem.* **29**, 1316–1331). For each subset, if the majority of ligands had measured $K_d$ values, all entries with $K_i$ were removed, and vice versa. Moreover, we only considered ligands with an unambiguous protonation state at neutral pH. $\mathrm{pK_a}$ values of ligands were predicted using Epik (Schrödinger). Ligands were excluded if they had two or more titratable groups with $\mathrm{pK_a}$ values between 6 to 8. These conditions resulted in a total of 16 protein subsets comprising 233 protein-ligand complexes, with the number of ligands per protein ranging from 5 to 39.

## Details of calculations for protein-ligand complexes

To find low energy conformers for each ligand in the free state, a conformational search protocol was performed using Macromodel (Schrödinger). Details of the search protocol are given in Supplementary Material. Solvent was represented by the generalized Born/solvent-accessible-surface-area (GBSA) implicit solvent model. The default OPLS_2005 force field was used for the conformational search, while the generalized AMBER force field (GAFF) was used throughout the rest of the calculation. Low-energy structures were clustered and the lowest-energy conformation was saved from each of the largest 30 clusters. The bound-state conformation from the crystal structure of the protein-ligand complex was manually added to the conformation set. Any previously-chosen conformations with an RMSD of less than 1.0 Å to the bound-state structure were discarded.

For the search for low-energy conformers of each ligand, we performed 5000 steps of systematic pseudo-Monte Carlo (SPMC) search with a 25 kcal/mol energy cutoff and 1.0 Å root mean square deviation (RMSD) cutoff. A conformation was saved only if its energy was within 25 kcal/mol of the lowest energy structure found and the structure is has an RMSD of at least 1.0 Å from all saved conformations. These cutoffs were used to prevent adopting high-energy or redundant conformations.

Generated structures were subjected to a local minimization with GAFF before clustering. Minimized conformations were clustered by distance matrix using the XCluster tool in MacroModel. For some rigid molecules, the conformational search generated fewer than 30 non-redundant clusters. For those cases all clusters were chosen.

Structures were used to initiate MD simulations and divided into wells, as described in the text. Wells with fewer than 1000 structures were discarded. The free energy calculations for wells with this few structures tended to be noisy and very high in free energy on average, so that they did not contribute to the overall free energy.

Free energy perturbation calculations were carried out between the quasiharmonic potential and the actual energy surface. All structure snapshots from the previous MD simulation were reused in this calculation. Here we simply performed one-step perturbation between the actual and QH potentials, using the

Zwanzig formula, although it would be possible to use more advanced perturbation methods such as the Bennett acceptance ratio method.

The binding pocket was taken to consist of all residues having at least one atom within 5.0 Å of any of the ligands in the protein subset. The part of the protein chosen was therefore the same for all ligands. Every peptide fragment in the pockets was capped with two extra protecting groups on each side: an acetyl starting group and a N-methylacetamide end group. All metals were removed from the receptor. The protonation state of each receptor was assigned at neutral pH using the program PDB2PQR. Ligand structures were adopted directly from PDBbind database and manually adjusted if needed (in several cases, bond-order information needed to be adjusted in the mol2 file from the PDBbind database, in order for the file to be successfully read by Maestro). The protonation state of the ligands were assigned using Epik (Schrödinger).

Most CPU time was spent on MD simulation of the ligands in the multiple wells representing the free state. The actual simulation time depended on the ligand size, the number of energy wells, and the length of the simulation in each well. For a 38-atom ligand with 30 wells, the calculation took $\sim$6 hours on a single AMD Opteron 246 processor. Note that all MD or MC simulations in each well are independent from each other, so it is trivial to perform these simulations in parallel.

We examined convergence with respect to the number of free energy wells used in the calculation, for the four largest ligands in the data set. For each molecule, multistate QH/FEP was applied using 20, 30, and 40 wells. For these ligands, the configuration integral was very similar for the calculations using 30 and 40 wells; the mean difference for the four ligands was 1.5%, and the largest difference was 3.5%, indicating that we could achieve a reasonable degree of convergence using 30 wells. Since convergence was examined for only the largest ligands, it is expected that these finite-sampling errors represent an upper bound for the remainder of the ligands.

## Rank correlation coefficients

In statistics, both $\tau$ and $\rho$ are used to measure the degree of correspondence between two rankings of the same set of items and to assess its significance. Both $\tau$ and $\rho$ range from $-1$ to 1. A value of 1 means perfect agreement between the two rankings, while a value of $-1$ means that experimental and calculated rankings place the compounds in opposite order. Here

$$\tau \equiv \frac{n_c - n_d}{n(n-1)/2} \tag{9}$$

and

$$\rho \equiv 1 - \frac{6\sum_{i=1}^{n}(R_i^{\mathrm{exp}} - R_i^{\mathrm{calc}})^2}{n(n^2-1)}. \tag{10}$$

In both equations, $n$ denotes the number of ligands in the subset. In equation 9, $n_c$ denotes the number of concordant pairs of ligands (pairs for which ligands appear in the same order in both rankings) and $n_d$ denotes the number of discordant pairs (pairs for which ligands appear in one order in one ranking, and the opposite order in the other ranking). In equation 10, $R_i^{\mathrm{exp}}$ is the rank of a ligand $i$ according to $\log K_d$ or $\log K_i$, while $R_i^{\mathrm{calc}}$ is the rank according to the calculated $\Delta G$.

Table S1: Protein-ligand complexes examined in this study. The last two columns give $-\log K_d$ or $-\log K_i$ for the highest- and lowest-affinity compounds in the protein subset.

| | Protein | PDB codes of complexes | Measurement | $pK_{max}$ | $pK_{min}$ |
|---|---|---|---|---|---|
| 1 | Serine/threonine-protein kinase Chk1 | 1nvq 1nvr 1nvs 2br1 2brb 2brm 2c3j 2c3l | $K_i$ | 8.25 | 4.86 |
| 2 | Acetylcholinesterase | 1e66 1gpk 1gpn 1h22 1h23 | $K_i$ | 9.89 | 5.37 |
| 3 | Tyrosine phosphatase 1B | 1g7f 1nny 1g7g 1nl9 1no6 1nz7 1ony 1pyn 1qxk 1xbo | $K_i$ | 7.66 | 4.20 |
| 4 | Beta-glucosidase A | 1oif 1w3j 2cbu 2ces 2cet 2j77 2j78 2j79 2j7b 2j7d 2j7e 2j7f 2j7g 2j7h | $K_d$ | 8.02 | 4.89 |
| 5 | Trypsin | 1qb6 1f0u 1qb9 1qbo 1qb1 1qbn 1f0t 1pph 1ppc 1c1r 1c5q 1c5s 1c5t 1tng 1tnh 1tni 1tnk 1tnl 1gi1 1gi4 1gj6 1o2h 1o2n 1o2o 1o2q 1o2w 1o2z 1o30 1o33 1o36 1o3d 1o3i 1o3j 1bjv 1bju 1g3b 1g3c 1ghz 1oyq | $K_i$ | 7.74 | 1.49 |
| 6 | Thrombin | 2bvr 2bvs 1g30 1k21 1k22 1sb1 1ghv 1ghw 1ghy 1mu6 1mu8 1c5n 1c5o 1tom 1oyt 1c4u 1c4v | $K_i$ | 10.80 | 3.49 |
| 7 | Coagulation factor Xa | 1mq6 1g2l 1mq5 1fjs 1xka 2p95 1ezq 1f0r 1f0s 1ksn 1lpg 1lpk 1lpz 1nfy 2boh 1nfu 1nfw 1nfx | $K_i$ | 11.15 | 8.52 |
| 8 | Urokinase-type plasminogen activator | 1gj8 1gjc 1o3p 1c5x 1c5y 1gi7 1gj7 1gja 1gjd | $K_i$ | 7.89 | 4.20 |
| 9 | Stromelysin-1 | 1b8y 1ciz 1sln 1usn 2d1o 2usn | $K_i$ | 7.85 | 6.51 |
| 10 | Thermolysin | 1qf0 1qf1 1tlp 1os0 1tmn 1qf2 1z9g 1zdp 2tmn 5tln 5tmn | $K_i$ | 8.04 | 3.42 |
| 11 | Penicillin amidohydrolase | 1ai4 1ai5 1ai7 1ajn 1ajp 1ajq | $K_i$ | 9.34 | 2.23 |
| 12 | Carbonic anhydrase II | 1bn1 1bn3 1bn4 1bnn 1bnq 1bnt 1bnu 1bnv 1bnw 1cnw 1cnx 1cny 1g1d 1g52 1g54 1if7 1if8 1ttm 1xpz 1xq0 | $K_d$ | 10.52 | 6.34 |
| 13 | Scytalone dehydratase | 3std 4std 5std 6std 7std | $K_i$ | 11.11 | 8.64 |
| 14 | HIV-1 protease | 1ajv 1ajx 1dif 1g2k 1g35 1gno 1hbv 1hih 1hos 1hpo 1hps 1hpv 1hpx 1hvh 1hvi 1hvj 1hvk 1hvl 1ohr 1w5v 1w5w 1w5x 1w5y 2aqu 2bpv 2bpy 2bqv 2cej 2cen 7upj | $K_i$ | 11.40 | 6.37 |
| 15 | Endothiapepsin | 5er1 5er2 1epo 4er1 4er2 2er6 | $K_d$ | 9.30 | 6.02 |
| 16 | Oligopeptide binding protein | 1b2h 1b4z 1b46 1b4h 2olb 1b3l 1qka 1b9j 1b5h 1b3h 1b58 1b3g 1b0h 1jeu 1b3f 1jev 1b1h 1b5i 1b32 1b05 1b52 1jet 1b40 1qkb 1b51 1b5j 1b6h 1b7h | $K_d$ | 8.02 | 4.54 |

Table S2: Parameter sets for water and united atom *n*-butane.

| | Stretch | | Bend | | Torsion | |
|---|---|---|---|---|---|---|
| | $k_a$ (kcal/mol Å$^{-2}$) | $l_0$ (Å) | $k_b$ (kcal/mol rad$^{-2}$) | $\theta_0$ (degrees) | $V_{n,\mathrm{dih}}$ (kcal/mol) | $n$ |
| water (CHARMM19) | 450.0 | 0.960 | 40.0 | 104.5 | N/ A | |
| water (modified) | 350.0 | 0.940 | 25.0 | 108.5 | N/A | |
| *n*-butane (CHARMM19) | 225 | 1.52 / 1.54 | 45 | 111.0 | 1.6 | 3 |
| *n*-butane (modified) | 100 | 1.30 / 1.30 | 20 | 100.0 | 0.6 | 3 |

Table S3: Calculation of $-k_{\mathrm{B}}T \ln Z$ (kcal/mol) using the analytical, quasiharmonic (QH), and quasiharmonic/free energy perturbation (QH/FEP) methods for water and *n*-butane, using CHARMM19 and the modified parameter set given in Table S1. All non-bond terms were turned off.

| | Analytical | QH | | | QH/FEP | | |
|---|---|---|---|---|---|---|---|
| | | zero order | second order | fourth order | zero order | second order | fourth order |
| water (CHARMM19) | 1.694 | 1.6997096 | 1.6997097 | 1.6997097 | 1.693 ±0.001 | 1.693±0.001 | 1.693±0.001 |
| water (modified) | 1.467 | 1.7976566 | 1.7976556 | 1.7976556 | 1.465±0.003 | 1.465±0.003 | 1.465±0.003 |
| *n*-butane (CHARMM19) | 1.838 | 1.1602633 | 1.1632136 | 1.1632066 | 1.837±0.002 | 1.839±0.002 | 1.839±0.002 |
| *n*-butane (modified) | 0.811 | 0.3538247 | 0.3595158 | 0.3594912 | 0.806±0.003 | 0.813±0.003 | 0.813±0.003 |

Table S4: Differences in $-RT \ln Z$ (kcal/mol) for water and *n*-butane when force-field parameters are modified from CHARMM19 to another set, as listed in Table S1, and for the alchemical change of ethane to methanol.

| | Analytical | Multistep FEP | QH | QH/FEP |
|---|---|---|---|---|
| water (CHARMM19 → modified) | 0.226 | 0.225±0.001 | −0.098 | 0.228±0.004 |
| n-butane (CHARMM19 → modified) | 1.027 | 1.021±0.003 | 0.804 | 1.026±0.005 |
| ethane → methanol | N/A | 2.881±0.012 | 3.162 | 2.892±0.012 |

Table S5: Standard deviation of the binding free energy, receptor ligand interaction energy, the ligand conformational free energy, as well as the experimental $RT \log K_i$ values for 5 successfully predicted ligand subsets. All values are in kcal/mol.

| Protein | Binding free energy | Receptor ligand interaction | Ligand conformational energy | $RT \log K_i$ |
|---|---|---|---|---|
| 1 | 11.23 | 10.33 | 1.32 | 0.80 |
| 2 | 6.53 | 9.28 | 9.30 | 1.00 |
| 8 | 2.79 | 3.23 | 1.96 | 0.69 |
| 9 | 11.20 | 10.55 | 4.49 | 0.32 |
| 10 | 9.16 | 14.36 | 7.48 | 0.49 |

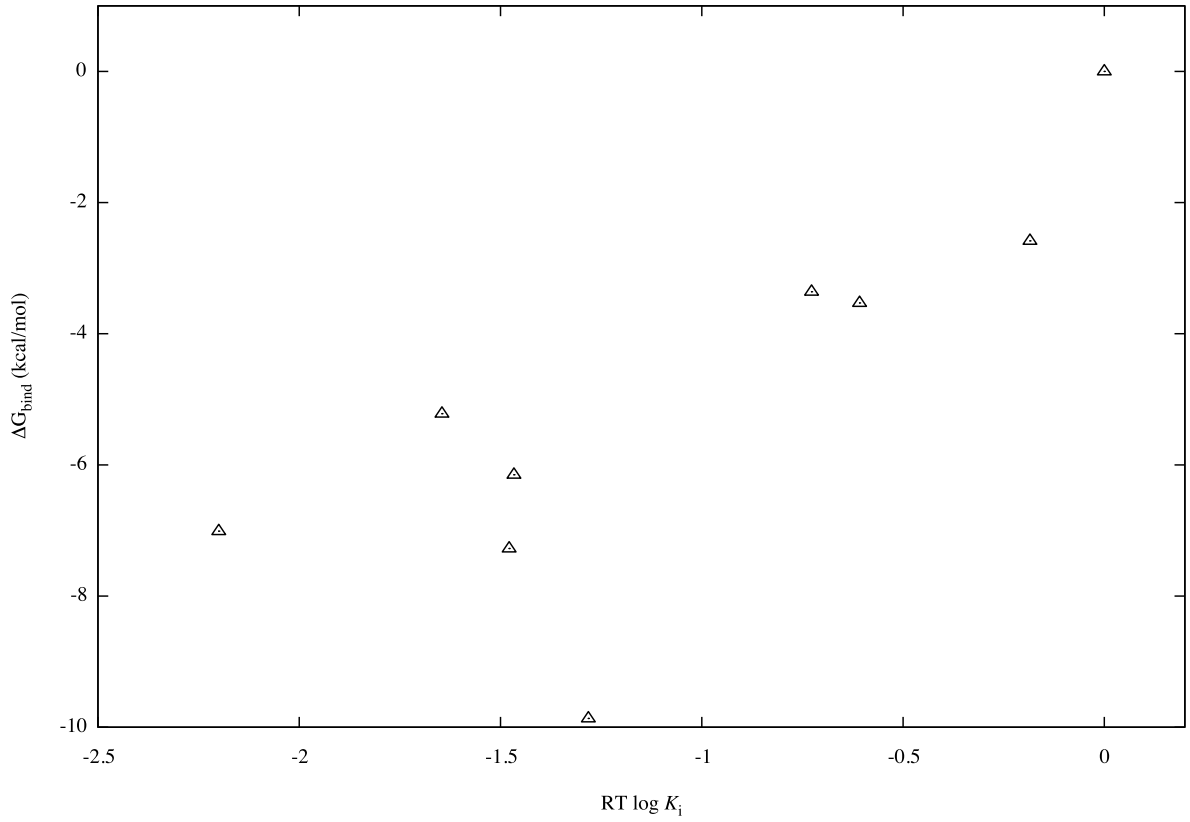


Figure S1: Experimental affinities $RT \log K_i$ versus $\Delta G_{\mathrm{bind}}$ (kcal/mol) for nine ligands bounds to urokinase-type plasminogen activator. A constant offset is added to both axes such that the weakest-bound ligand has a value of zero.

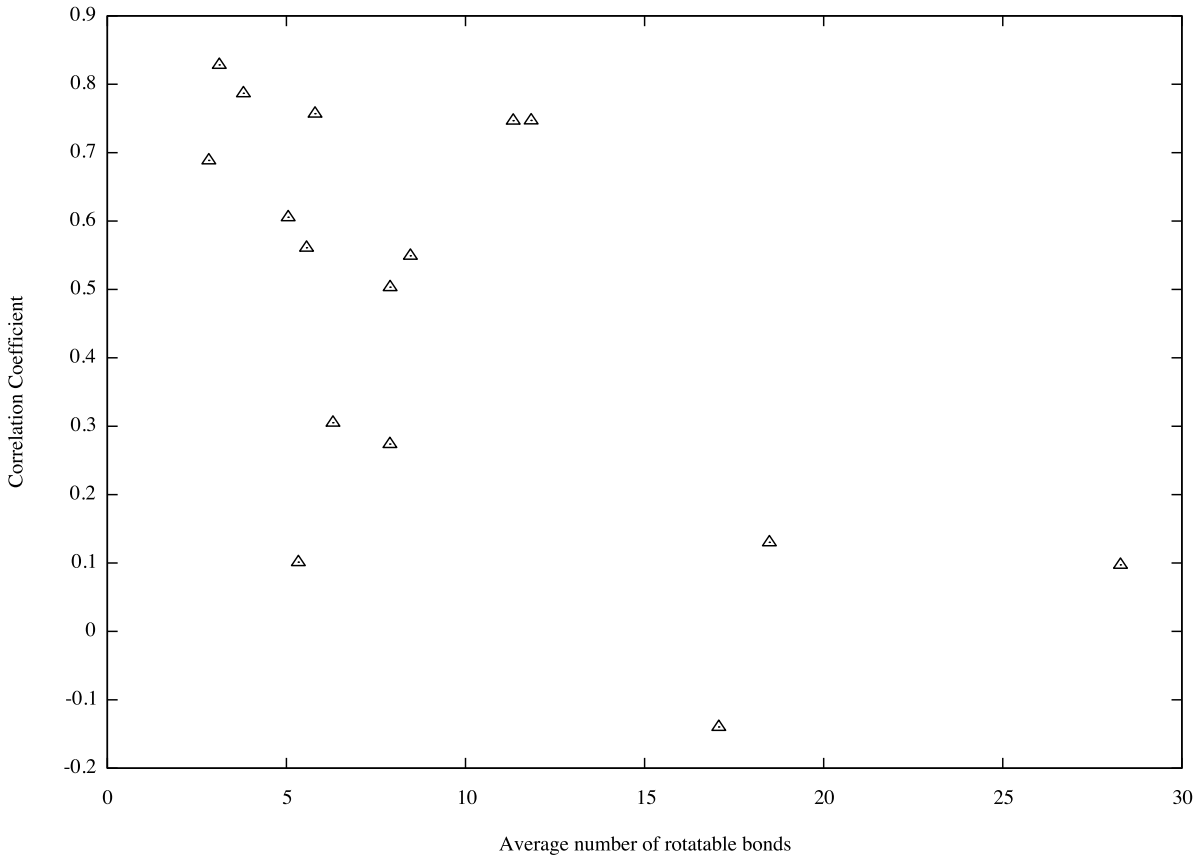


Figure S2: Correlation coefficients between $\Delta G_{\mathrm{bind}}$ and $\log K_i$ or $\log K_d$, as a function of the average number of rotatable bonds for all ligands in a protein subset, for each of the 16 proteins examined.

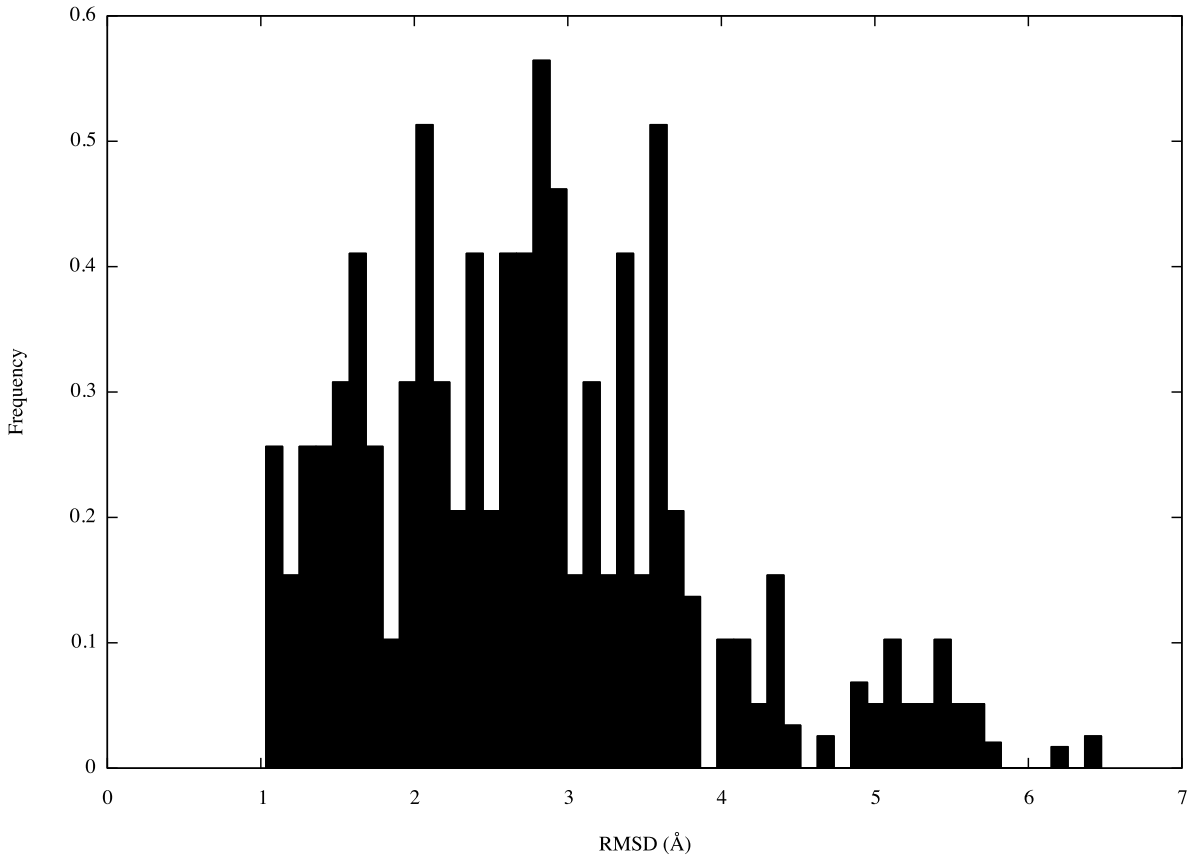


Figure S3: Histogram of the RMSD between the bound-state ligand conformation and the most favorable free-state conformation (i.e., reference conformation for the well with the largest calculated configuration integral).

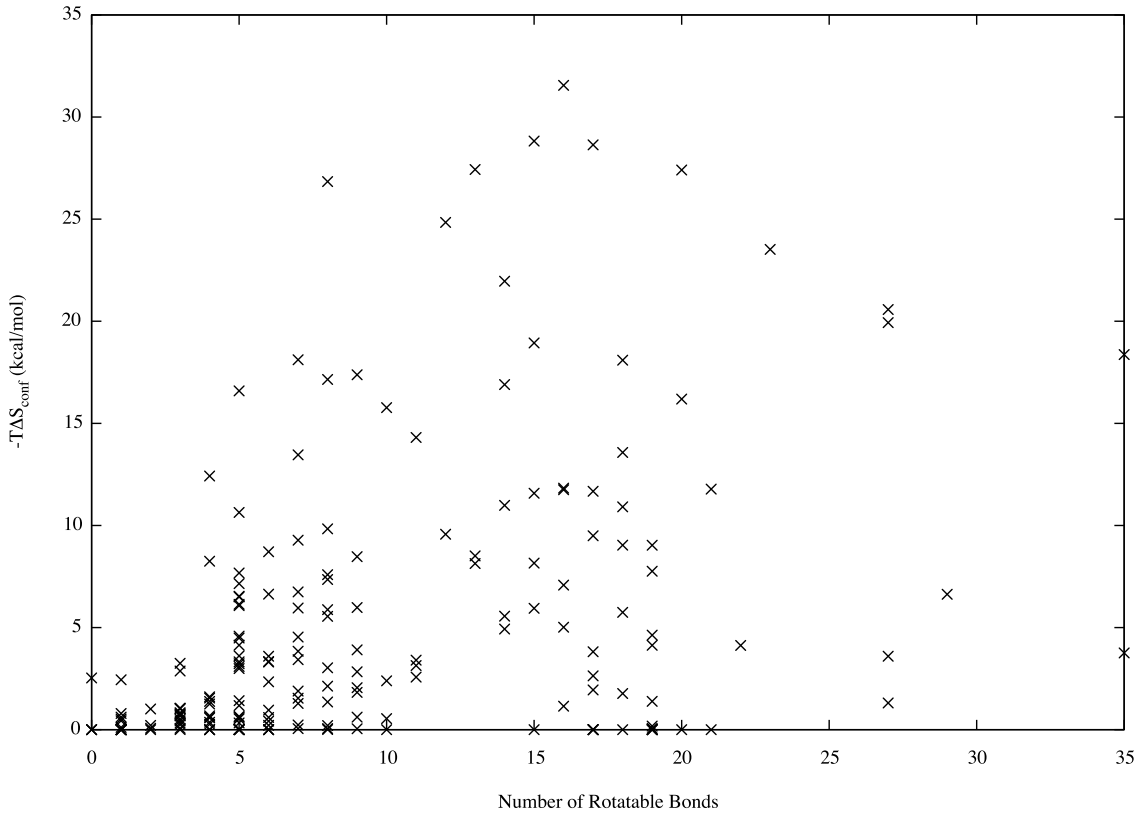


Figure S4: Entropic contribution to the ligand conformational free energy change upon binding, $-T\Delta S_{\mathrm{conf}}$, as a function of the number of rotatable bonds.