Supplemental Information to

Computing the binding affinity of a ligand buried deep inside a protein with the hybrid steered molecular dynamics

Oscar D Villarreal, Lili Yu, Roberto A Rodriguez and Liao Y Chen

In the first section, we present the detailed derivation of the central formula that relates the binding affinity to the potential of mean force (PMF) along the open-pull-close path and the partial partitions. Then we present figures for all the relevant data.

THEORETICAL DERIVATION

The absolute binding energy from the 3(m+n)-D PMF. Following the standard literature,[1, 2] the binding affinity at one binding site is

$$\frac{1}{k_D / c_0} = \frac{c_0 \int_{\text{site}} d^3 x_1^L \exp\left[-W[\mathbf{r}_1^L]/k_B T\right]}{\exp\left[-W[\mathbf{r}_{1\infty}^L]/k_B T\right]_{\text{bulk}}}. \tag{S1}$$

Here c_0 is the standard concentration. For clarity and for convenience of unit conversion, we use two different but equivalent forms, $c_0 = 1\,M$ on the left hand side and $c_0 = 6.02\times10^{-4}/\mathring{A}^3$ on the right hand side of the equation. k_B is the Boltzmann constant and T is the absolute temperature. The three-dimensional (3D) integrations are over the x-, y-, and z-coordinates of the ligand's position $\mathbf{r}_1^L$ that can be chosen as the center of mass of one segment or of the whole ligand. The integral has the units of $Å^3$ that renders the right hand side dimensionless as it should be. $W[\mathbf{r}_1^L]$ is the 3D PMF. The subscripts "site" and "bulk" indicate that $\mathbf{r}_1^L$ is near the PMF minimum and $\mathbf{r}_{1\infty}^L$ is in the bulk region far away from the protein, respectively.

Note that in the relationship between the 3-D and 3(m+n)-D PMFs,

$$\exp\left[-W[\mathbf{r}_1^L]/k_BT\right] = C\int \prod_{i=2}^{m} d^3x_i^L \prod_{i=1}^{n} d^3x_i^P \exp\left[-W[\mathbf{r}_1^L,\mathbf{r}_2^L,...,\mathbf{r}_m^L;\mathbf{r}_1^P,\mathbf{r}_2^P,...,\mathbf{r}_n^P]/k_BT\right], \quad \text{(S2)}$$

the 3(m+n-1)-D integration are over the (m+n-1) positions: $\left(\mathbf{r}_2^L,...,\mathbf{r}_m^L\right)$ of the m-1 pulling centers on the ligand and $\left(\mathbf{r}_1^P,...,\mathbf{r}_n^P\right)$ of the n pulling centers on the protein. *C* is the normalization constant that will be cancelled out in the following expressions.

Using the relationship given by Eq. (S2) in Eq. (S1) twice (once in the numerator for the binding site and once in the denominator for the bulk), one has the following expression for the binding affinity,

$$\frac{c_0}{k_D} = \frac{c_0\int_{\text{site}} \prod_{i=1}^{m} d^3x_i^L \prod_{i=1}^{n} d^3x_i^{\text{P}} \exp\left[-W[\mathbf{r}_1^L,\mathbf{r}_2^L,...,\mathbf{r}_m^L;\mathbf{r}_1^P,\mathbf{r}_2^P,...,\mathbf{r}_n^P]/k_BT\right]}{\int_{\text{bulk}} \prod_{i=2}^{m} d^3x_i^L \prod_{i=1}^{n} d^3x_i^{\text{P}} \exp\left[-W[\mathbf{r}_{1\infty}^L,\mathbf{r}_2^L,...,\mathbf{r}_m^L;\mathbf{r}_1^P,\mathbf{r}_2^P,...,\mathbf{r}_n^P]/k_BT\right]}. \quad \text{(S3)}$$

Now inserting the Boltzmann factor at a single state $\left(\mathbf{r}_{10}^L,\mathbf{r}_{20}^L,...,\mathbf{r}_{m0}^L;\mathbf{r}_{10}^P,\mathbf{r}_{20}^P,...,\mathbf{r}_{n0}^P\right)$ chosen from the bound state ensemble and the Boltzmann factor at the corresponding dissociated state $\left(\mathbf{r}_{1\infty}^L,\mathbf{r}_{2\infty}^L,...,\mathbf{r}_{m\infty}^L;\mathbf{r}_{10}^P,\mathbf{r}_{20}^P,...,\mathbf{r}_{n0}^P\right)$, the binding affinity can be expressed as three contributing factors: The partial partition function at the binding site Z_{m+n0} of the ligand-protein complex, the PMF difference between two chosen states $\left(\mathbf{r}_{10}^L,\mathbf{r}_{20}^L,...,\mathbf{r}_{m0}^L;\mathbf{r}_{10}^P,\mathbf{r}_{20}^P,...,\mathbf{r}_{n0}^P\right)$ and $\left(\mathbf{r}_{1\infty}^L,\mathbf{r}_{2\infty}^L,...,\mathbf{r}_{m\infty}^L;\mathbf{r}_{10}^P,\mathbf{r}_{20}^P,...,\mathbf{r}_{n0}^P\right)$, and the partial partition function in the dissociated state $Z_{m-1+n\infty}$. Mathematically,

$$\frac{c_0}{k_D}=\frac{c_0\int_{\text{site}}\prod_{i=1}^{m}d^3x_i^L\prod_{i=1}^{n}d^3x_i^P\exp\left[-W[\mathbf{r}_1^L,\mathbf{r}_2^L,...,\mathbf{r}_m^L;\mathbf{r}_1^P,\mathbf{r}_2^P,...,\mathbf{r}_n^P]/k_BT\right]}{\exp\left[-W[\mathbf{r}_{10}^L,\mathbf{r}_{20}^L,...,\mathbf{r}_{m0}^L;\mathbf{r}_{10}^P,\mathbf{r}_{20}^P,...,\mathbf{r}_{n0}^P]/k_BT\right]}\times$$
$$\frac{\exp\left[-W[\mathbf{r}_{10}^L,\mathbf{r}_{20}^L,...,\mathbf{r}_{m0}^L;\mathbf{r}_{10}^P,\mathbf{r}_{20}^P,...,\mathbf{r}_{n0}^P]/k_BT\right]}{\exp\left[-W[\mathbf{r}_{1\infty}^L,\mathbf{r}_{2\infty}^L,...,\mathbf{r}_{m\infty}^L;\mathbf{r}_{10}^P,\mathbf{r}_{20}^P,...,\mathbf{r}_{n0}^P]/k_BT\right]}\times$$
$$\frac{\exp\left[-W[\mathbf{r}_{1\infty}^L,\mathbf{r}_{2\infty}^L,...,\mathbf{r}_{m\infty}^L;\mathbf{r}_{10}^P,\mathbf{r}_{20}^P,...,\mathbf{r}_{n0}^P]/k_BT\right]}{\int_{\text{bulk}}\prod_{i=2}^{m}d^3x_i^L\prod_{i=1}^{n}d^3x_i^P\exp\left[-W[\mathbf{r}_{1\infty}^L,\mathbf{r}_2^L,...,\mathbf{r}_m^L;\mathbf{r}_1^P,\mathbf{r}_2^P,...,\mathbf{r}_n^P]/k_BT\right]}$$
$$=\frac{c_0Z_{m+n0}}{Z_{m-1+n\infty}}\exp\left[\frac{-\Delta W_{0,\infty}}{k_BT}\right].$$

(S4)

Here the subscript 0 refers to the bound/holo state and ∞ refers to the dissociated/apo state. The one apo state $\left(\mathbf{r}_{1\infty}^L,\mathbf{r}_{2\infty}^L,...,\mathbf{r}_{m\infty}^L;\mathbf{r}_{10}^P,\mathbf{r}_{20}^P,...,\mathbf{r}_{n0}^P\right)$ can be connected to the one holo state $\left(\mathbf{r}_{10}^L,\mathbf{r}_{20}^L,...,\mathbf{r}_{m0}^L;\mathbf{r}_{10}^P,\mathbf{r}_{20}^P,...,\mathbf{r}_{n0}^P\right)$ via many curves in the 3(m+n)-D space but the PMF is a function of state; thus the computation of the PMF difference between the two states can be achieved along a single curve passing through them both. Again, the 3(m+n)-D PMF difference

$$\Delta W_{0,\infty}=W[\mathbf{r}_{10}^L,\mathbf{r}_{20}^L,...,\mathbf{r}_{m0}^L;\mathbf{r}_{10}^P,\mathbf{r}_{20}^P,...,\mathbf{r}_{n0}^P]-W[\mathbf{r}_{1\infty}^L,\mathbf{r}_{2\infty}^L,...,\mathbf{r}_{m\infty}^L;\mathbf{r}_{10}^P,\mathbf{r}_{20}^P,...,\mathbf{r}_{n0}^P] \quad \text{(S5)}$$

is between one chosen bound state and its corresponding dissociated state. This PMF difference can be computed not by subtracting one PMF value from another PMF value but by means of the SMD simulations described in the subsection immediately following this current subsection. Note that the one chosen position of the m+n pulling centers in the bound state,

$$\left(\mathbf{r}_{10}^L,...,\mathbf{r}_{m0}^L;\mathbf{r}_{10}^P,...,\mathbf{r}_{n0}^P\right)=(x_{10}^L,y_{10}^L,z_{10}^L,...,x_{m0}^L,y_{m0}^L,z_{m0}^L;x_{10}^P,y_{10}^P,z_{10}^P,...,x_{n0}^P,y_{n0}^P,z_{n0}^P) \quad \text{(S6)}$$

is the starting point for SMD runs. It is taken from the bound state ensemble of the system. It does not have to be the minimum of the PMF but any one state in its close neighborhood. Note that we take the collection

of coordinate vectors, e.g., Eq. (S6), as a single-row $1\times 3(m+n)$ matrix. The one state chosen from the dissociated state ensemble

$$\left(\mathbf{r}_{1\infty}^{L},...,\mathbf{r}_{m\infty}^{L};\mathbf{r}_{10}^{P},...,\mathbf{r}_{n0}^{P}\right)=(x_{1\infty}^{L},y_{1\infty}^{L},z_{1\infty}^{L},...,x_{m\infty}^{L},y_{m\infty}^{L},z_{m\infty}^{L};x_{10}^{P},y_{10}^{P},z_{10}^{P},...,x_{n0}^{P},y_{n0}^{P},z_{n0}^{P}) \tag{S7}$$

is related to the SMD starting point by a large enough displacement in the 3(m+n)-D space,

$$(\mathbf{r}_{1\infty}^{L},...,\mathbf{r}_{m\infty}^{L};\mathbf{r}_{10}^{P},...,\mathbf{r}_{n0}^{P})=(\mathbf{r}_{10}^{L}+\mathbf{v}_{\mathrm{d}}t,...,\mathbf{r}_{m0}^{L}+\mathbf{v}_{\mathrm{d}}t;\mathbf{r}_{10}^{P},...,\mathbf{r}_{n0}^{P}). \tag{S8}$$

Here $\mathbf{v}_{\mathrm{d}}$ is the constant velocity of the SMD pulling and t is the time it takes to steer/pull the ligand from the bound state to the dissociated state. However, it should be emphasized that the path from the bound state $\left(\mathbf{r}_{10}^{L},\mathbf{r}_{20}^{L},...,\mathbf{r}_{m0}^{L};\mathbf{r}_{10}^{P},\mathbf{r}_{20}^{P},...,\mathbf{r}_{n0}^{P}\right)$ to the dissociated state $\left(\mathbf{r}_{1\infty}^{L},\mathbf{r}_{2\infty}^{L},...,\mathbf{r}_{m\infty}^{L};\mathbf{r}_{10}^{P},\mathbf{r}_{20}^{P},...,\mathbf{r}_{n0}^{P}\right)$ is not a straight line in the 3(m+n)-D space as Eq. (S8) might suggest. Instead, the dissociation path consists of three straight lines: First, from $\left(\mathbf{r}_{10}^{L},\mathbf{r}_{20}^{L},...,\mathbf{r}_{m0}^{L};\mathbf{r}_{10}^{P},\mathbf{r}_{20}^{P},...,\mathbf{r}_{n0}^{P}\right)$ to $\left(\mathbf{r}_{10}^{L},\mathbf{r}_{20}^{L},...,\mathbf{r}_{m0}^{L};\mathbf{r}_{10}^{P}+\mathbf{v}_{1}^{P}\Delta t,\mathbf{r}_{20}^{P}+\mathbf{v}_{2}^{P}\Delta t,...,\mathbf{r}_{n0}^{P}+\mathbf{v}_{n}^{P}\Delta t\right)$, pulling the n centers on the protein to open the binding cavity up so that the ligand is exposed for extraction. Second, from there to $\left(\mathbf{r}_{1\infty}^{L},\mathbf{r}_{2\infty}^{L},...,\mathbf{r}_{m\infty}^{L};\mathbf{r}_{10}^{P}+\mathbf{v}_{1}^{P}\Delta t,\mathbf{r}_{20}^{P}+\mathbf{v}_{2}^{P}\Delta t,...,\mathbf{r}_{n0}^{P}+\mathbf{v}_{n}^{P}\Delta t\right)$, pulling the ligand out of the cavity to the bulk region far from the protein while holding the centers on the protein all fixed. Third, from there to $\left(\mathbf{r}_{1\infty}^{L},\mathbf{r}_{2\infty}^{L},...,\mathbf{r}_{m\infty}^{L};\mathbf{r}_{10}^{P},\mathbf{r}_{20}^{P},...,\mathbf{r}_{n0}^{P}\right)$, pulling the centers on the protein, in the direction exactly opposite the first straight line, back to their initial positions while holding the ligand fixed in the bulk region far away from the protein. It is pointed here that the third part is preferably implemented not as closing back the binding site directly but as the inverse of opening up the cavity from $\left(\mathbf{r}_{1\infty}^{L},\mathbf{r}_{2\infty}^{L},...,\mathbf{r}_{m\infty}^{L};\mathbf{r}_{10}^{P},\mathbf{r}_{20}^{P},...,\mathbf{r}_{n0}^{P}\right)$ to $\left(\mathbf{r}_{1\infty}^{L},\mathbf{r}_{2\infty}^{L},...,\mathbf{r}_{m\infty}^{L};\mathbf{r}_{10}^{P}+\mathbf{v}_{1}^{P}\Delta t,\mathbf{r}_{20}^{P}+\mathbf{v}_{2}^{P}\Delta t,...,\mathbf{r}_{n0}^{P}+\mathbf{v}_{n}^{P}\Delta t\right)$ after an equilibrium MD run of the apo protein with all the pulling centers fixed at $\left(\mathbf{r}_{1\infty}^{L},\mathbf{r}_{2\infty}^{L},...,\mathbf{r}_{m\infty}^{L};\mathbf{r}_{10}^{P},\mathbf{r}_{20}^{P},...,\mathbf{r}_{n0}^{P}\right)$.

Note that pulling velocities of the centers on the protein need to be chosen appropriately to open up the binding cavity. The final result of binding affinity does not depend on the choice of these velocities but the computational efficiency does. In all the five systems studied in this work, we chose to pull the alpha carbons of the more flexible parts of the protein in different directions to expose the ligand in the easiest manner simply by visual inspections of the binding complexes. It turned out that our first choices for all systems are valid and good for the computation of the binding affinities.

The partial partition function Z_{m+n0} of the bound state has the integration over m+n centers and thus has the units of $Å^{3(m+n)}$,

$$Z_{m+n0} = \int_{\text{site}} \prod_{i=1}^{m} d^3x_i^L \prod_{i=1}^{n} d^3x_i^P \exp\left[-\begin{pmatrix} W[\mathbf{r}_1^L, \mathbf{r}_2^L, ..., \mathbf{r}_m^L; \mathbf{r}_1^P, \mathbf{r}_2^P, ..., \mathbf{r}_n^P] - \\ W[\mathbf{r}_{10}^L, \mathbf{r}_{20}^L, ..., \mathbf{r}_{m0}^L; \mathbf{r}_{10}^P, \mathbf{r}_{20}^P, ..., \mathbf{r}_{n0}^P] \end{pmatrix} / k_B T \right]. \tag{S9}$$

Note that, in the bound state, the m+n pulling centers all fluctuate around the one initial state $\left(\mathbf{r}_{10}^L, \mathbf{r}_{20}^L, ..., \mathbf{r}_{m0}^L; \mathbf{r}_{10}^P, \mathbf{r}_{20}^P, ..., \mathbf{r}_{n0}^P\right)$ chosen from the bound state ensemble. When the binding is tight, one can approximate the fluctuations as Gaussian in the neighborhood of the PMF minimum. The coordinates of the minimum of a Gaussian distribution are equal to the average coordinates, of course, $\left(\left\langle \mathbf{r}_1^L \right\rangle, \left\langle \mathbf{r}_2^L \right\rangle, ..., \left\langle \mathbf{r}_m^L \right\rangle; \left\langle \mathbf{r}_1^P \right\rangle, \left\langle \mathbf{r}_2^P \right\rangle, ..., \left\langle \mathbf{r}_n^P \right\rangle\right)$.

$$Z_{m+n0} = (2\pi)^{3(m+n)/2} Det^{1/2}(\Sigma_{m+no}) \exp\left[\Delta_{m+no} / k_B T\right]. \tag{S10}$$

Here the dimensionless quantity $\Delta_{m+no}/k_B T$ gives a measure of how far $\left(\mathbf{r}_{10}^L, \mathbf{r}_{20}^L, ..., \mathbf{r}_{m0}^L; \mathbf{r}_{10}^P, \mathbf{r}_{20}^P, ..., \mathbf{r}_{n0}^P\right)$, the initial state chosen for SMD, is from the PMF minimum $\left(\left\langle \mathbf{r}_1^L \right\rangle, \left\langle \mathbf{r}_2^L \right\rangle, ..., \left\langle \mathbf{r}_m^L \right\rangle; \left\langle \mathbf{r}_1^P \right\rangle, \left\langle \mathbf{r}_2^P \right\rangle, ..., \left\langle \mathbf{r}_n^P \right\rangle\right)$,

$$\begin{aligned} \Delta_{m+no}/k_B T = \tfrac{1}{2}\left(\left\langle \mathbf{r}_1^L \right\rangle - \mathbf{r}_{10}^L, ..., \left\langle \mathbf{r}_m^L \right\rangle - \mathbf{r}_{m0}^L; \left\langle \mathbf{r}_1^P \right\rangle - \mathbf{r}_{10}^P, ..., \left\langle \mathbf{r}_n^P \right\rangle - \mathbf{r}_{n0}^P\right) \Sigma_{m+no}^{-1} \times \\ \left(\left\langle \mathbf{r}_1^L \right\rangle - \mathbf{r}_{10}^L, ..., \left\langle \mathbf{r}_m^L \right\rangle - \mathbf{r}_{m0}^L; \left\langle \mathbf{r}_1^P \right\rangle - \mathbf{r}_{10}^P, ..., \left\langle \mathbf{r}_n^P \right\rangle - \mathbf{r}_{n0}^P\right)^T. \end{aligned} \tag{S11}$$

Σ_{m+no} is the 3(m+n)×3(m+n) matrix of the fluctuations/deviations of the pulling center coordinates $\delta x_1^L = x_1^L - \left\langle x_1^L \right\rangle$ *etc.* The superscript T indicates the transpose of the matrix.

$$\Sigma_{m+no} = \left\langle \begin{array}{c} \left(\left\langle \mathbf{r}_1^L \right\rangle - \mathbf{r}_{10}^L, ..., \left\langle \mathbf{r}_m^L \right\rangle - \mathbf{r}_{m0}^L; \left\langle \mathbf{r}_1^P \right\rangle - \mathbf{r}_{10}^P, ..., \left\langle \mathbf{r}_n^P \right\rangle - \mathbf{r}_{n0}^P \right)^T \times \\ \left(\left\langle \mathbf{r}_1^L \right\rangle - \mathbf{r}_{10}^L, ..., \left\langle \mathbf{r}_m^L \right\rangle - \mathbf{r}_{m0}^L; \left\langle \mathbf{r}_1^P \right\rangle - \mathbf{r}_{10}^P, ..., \left\langle \mathbf{r}_n^P \right\rangle - \mathbf{r}_{n0}^P \right) \end{array} \right\rangle . \quad \text{(S12)}$$

Σ_{m+no}^{-1} is the inverse matrix of Σ_{m+no} which can be accurately evaluated by running equilibrium MD in the associated state of the ligand-protein complex.

The partial partition of the apo state, $Z_{m-1+n\infty}$, refers to the fact that the m-1 centers on the ligand and the n pulling centers on the protein are free to fluctuate around the one final state $\left(\mathbf{r}_{1\infty}^L, \mathbf{r}_{2\infty}^L, ..., \mathbf{r}_{m\infty}^L; \mathbf{r}_{10}^P, \mathbf{r}_{20}^P, ..., \mathbf{r}_{n0}^P \right)$ in which $\mathbf{r}_1^L = \mathbf{r}_{1\infty}^L$ is fixed. Thus the integration is over m+n-1 variables,

$$Z_{m-1+n\infty} = \int_{\text{bulk}} \prod_{i=2}^{m} d^3 x_i^L \prod_{i=1}^{n} d^3 x_i^P \exp\left[-\left(\begin{array}{l} W[\mathbf{r}_{1\infty}^L, \mathbf{r}_2^L, ..., \mathbf{r}_m^L; \mathbf{r}_1^P, \mathbf{r}_2^P, ..., \mathbf{r}_n^P] - \\ W[\mathbf{r}_{1\infty}^L, \mathbf{r}_{2\infty}^L, ..., \mathbf{r}_{m\infty}^L; \mathbf{r}_{10}^P, \mathbf{r}_{20}^P, ..., \mathbf{r}_{n0}^P] \end{array} \right) / k_B T \right] . \quad \text{(S13)}$$

This partial partition function has the units of $\text{Å}^{3(m+n-1)}$. It should be noted that, in the dissociated state, the ligand and the protein are not interacting with each other. Therefore, the partial partition function can be factored into two separate partial partition functions of the ligand and the protein respectively, $Z_{m-1+n\infty} = Z_{m-1\infty}^L Z_{n\infty}^P$. $Z_{n\infty}^P$ contains the fluctuations of the n pulling centers on the protein in the dissociated state. The Gaussian approximation of this partial partition for the n centers on the protein is identical to that of the bound state partial partition with m+n now being replaced with n. The partial partition for the pulling centers on the ligand in the dissociated state, $Z_{m-1\infty}^L$, contains the fluctuations of the m-1 pulling centers on the ligand in the dissociated state while one of the centers is fixed. Its computation was fully described in Ref.[3] for an arbitrary integer m>0. In particular, in this paper, we use only m=1 or m=2. In the case of m=1, $Z_{1-1\infty}^L = 1$. In the case of m=2, we have

$$
\begin{aligned}
Z_{2-1\infty}^{L} &= \int_{bulk} d^3x_2^L \exp\left[-\left(W[\mathbf{r}_{1\infty}^L, \mathbf{r}_2^L] - W[\mathbf{r}_{1\infty}^L, \mathbf{r}_{2\infty}^L]\right)/k_BT\right] \\
&= 4\pi \int dr \exp\left[-\left(W[r] - W[r_\infty]\right)/k_BT\right] r^2 \\
&= 4\pi\left[\langle r\rangle^2 + \langle \delta r^2\rangle\right] \exp\left[-\frac{\left(r_\infty - \langle r\rangle\right)^2}{2\langle \delta r^2\rangle}\right] \sqrt{2\pi\langle \delta r^2\rangle}\,.
\end{aligned}
\tag{S14}
$$

Here $r = \left|\mathbf{r_2} - \mathbf{r_{1\infty}}\right|$ is the distance between the two pulling centers on the ligand. Note that $r_\infty = \left|\mathbf{r}_{2\infty}^L - \mathbf{r}_{1\infty}^L\right| = \left|\mathbf{r}_{20}^L - \mathbf{r}_{10}^L\right|$ because all the centers on the ligand are pulled with one speed. $W[r]$, a function of r, is the 1D PMF in the dissociated state for stretching the ligand between the two pulling centers. It can be evaluated, as in the second line of the equation (S14), by conducting SMD runs of steering the second pulling center $\mathbf{r_2}$ to and from the first pulling center that is fixed at $\mathbf{r_{1\infty}}$ along the axis passing through $(\mathbf{r_{1\infty}}, \mathbf{r_{2\infty}})$. It can also be computed in the Gaussian approximation as the third line of the equation (S14) which is valid when the ligand, upon binding inside the protein, is not significantly stretched or compressed between the two chosen pulling centers. $\langle r\rangle$ is the mean distance between the two pulling centers on the ligand and $\delta r = r - \langle r\rangle$ represents the fluctuation. The mean square deviation $\langle \delta r^2\rangle$ gives a measure of the ligand's flexibility in the apo state which is generally different from that of the holo state. Note that, in the Gaussian approximation,

$$
W[r] \approx W[\langle r\rangle] + \frac{1}{2}W''[\langle r\rangle](r - \langle r\rangle)^2 = W[\langle r\rangle] + \frac{1}{2}\delta r^2 \Big/ \langle \delta r^2\rangle, \tag{S15}
$$

the mean $\langle r\rangle$ is also the minimum of the 1D PMF. The factor of 4π in Eq. (S14) comes from the angular integration because the ligand in the apo state is free to rotate around one pulling center.

Again, the use of $c_0 = 6.02\times10^{-4}\,/\,\mathring{A}^3$ on the right hand side of Eq. (S3) renders it a pure number as desired. And the dissociation constant will conveniently be in the unit of M=mol/L. In summary, we have the following formulas for the binding affinity and the absolute binding free energy, respectively,

$$\frac{c_0}{k_D} = \frac{c_0 Z_{m+n0}}{Z^L_{m-1\infty} Z^P_{n\infty}} \exp\left[-\frac{\Delta W_{0,\infty}}{k_B T}\right], \quad \Delta G_{\text{binding}} = k_B T \ln\left[\frac{Z^L_{m-1\infty} Z^P_{n\infty}}{Z_{m+n0} c_0}\right] + \Delta W_{0,\infty}\,, \tag{S16}$$

once we have computed the PMF difference along the open-pull-close path and the three partial partitions. When the conformation changes between the apo and the holo protein are small such as in the four complexes of this study, the Gaussian approximation is valid and accurate for the partial partition $Z^P_{n\infty}$ characterizing the fluctuations of the n pulling centers on the apo protein. Otherwise, other ways need to be explored to better approximate $Z^P_{n\infty}$.

PMF from SMD simulations. In an SMD[4] simulation of the current literature, one steers/pulls one center of mass of one selection of atoms, using a spring with a carefully chosen stiffness (spring constant). The use of a spring of finite stiffness introduces additional fluctuation and dissipation in the added degrees of freedom.[5] In this paper, we choose m+n segments (mutually exclusive m+n selections of atoms) of the ligand and the protein for steering/pulling with m+n infinitely stiff springs (m, n=1, 2, 3.....). Namely, the m+n centers of mass of the chosen m+n segments will be controlled as functions of time t

$$\mathbf{r}_i = \mathbf{r}_{iA} \pm \mathbf{v}_{\mathbf{d}i} t, \;\; i = 1...m+n \tag{S17}$$

while all the other degrees of freedom of the system are freely subject to stochastic dynamics. Here $\mathbf{r}_i = (x_i, y_i, z_i)$ is the center of mass coordinates of the *i*-th segments, $\mathbf{v}_{\mathbf{d}i}$ is the pulling velocity, and $\mathbf{r}_{iA}$ are coordinates of the centers of mass of the steered segments at the end state A. The + and – signs are for the forward and reverse pulling paths, respectively. $\{\mathbf{r}_i\}$ denotes $\left(\mathbf{r}^L_1, \mathbf{r}^L_2, ..., \mathbf{r}^L_m; \mathbf{r}^{\mathrm{P}}_1, \mathbf{r}^{\mathrm{P}}_2, ..., \mathbf{r}^{\mathrm{P}}_n\right)$ etc. We adopt the multi-sectional scheme of Ref. [6]. The path from the bound state to the dissociated state is divided

into a number of sections. Within a given section whose end states are marked as A and B, respectively, multiple forward and reverse pulling paths are sampled along which the work done to the system is recorded. The Gibbs free-energy difference (namely, the PMF or the reversible work) is computed via the Brownian-dynamics fluctuation-dissipation theorem (BD-FDT)[7] as follows:

$$W\left[\left\{\mathbf{r}_i\right\}\right]-W\left[\left\{\mathbf{r}_{iA}\right\}\right]=-k_BT\ln\left(\frac{\left\langle\exp[-W_{\{\mathbf{r}_{iA}\}\to\{\mathbf{r}_i\}}/2k_BT]\right\rangle_F}{\left\langle\exp[-W_{\{\mathbf{r}_i\}\to\{\mathbf{r}_{iA}\}}/2k_BT]\right\rangle_R}\right). \tag{S18}$$

Here the brackets with subscript F/R represent the statistical average over the forward/reverse paths. $W_{\{\mathbf{r}_{iA}\}\to\{\mathbf{r}_i\}}$ is the work done to the system along a forward path when the pulling centers are steered from A to **r**. $W_{\{\mathbf{r}_i\}\to\{\mathbf{r}_{iA}\}}=W_{\{\mathbf{r}_{iB}\}\to\{\mathbf{r}_{iA}\}}-W_{\{\mathbf{r}_{iB}\}\to\{\mathbf{r}_i\}}$ is the work for the part of a reverse path when the centers are pulled from **r** to A. $\left\{\mathbf{r}_{iA}\right\}$, $\left\{\mathbf{r}_i\right\}$, and $\left\{\mathbf{r}_{iB}\right\}$ are the coordinates of the centers of mass of the steered segments at the end state A, the general state **r**, and the end state B of the system, respectively. At each end of a section, A/B, the system is equilibrated for a time long enough to reach conditioned equilibrium while the steered centers are fixed at $\left\{\mathbf{r}_{iA}\right\}/\left\{\mathbf{r}_{iB}\right\}$. In this way, running SMD section by section, we map the PMF $W\left[\left\{\mathbf{r}_i\right\}\right]$ as a function of the steered centers along a chosen path from the bound state to the dissociated state.

SUPPLEMENTAL FIGURES

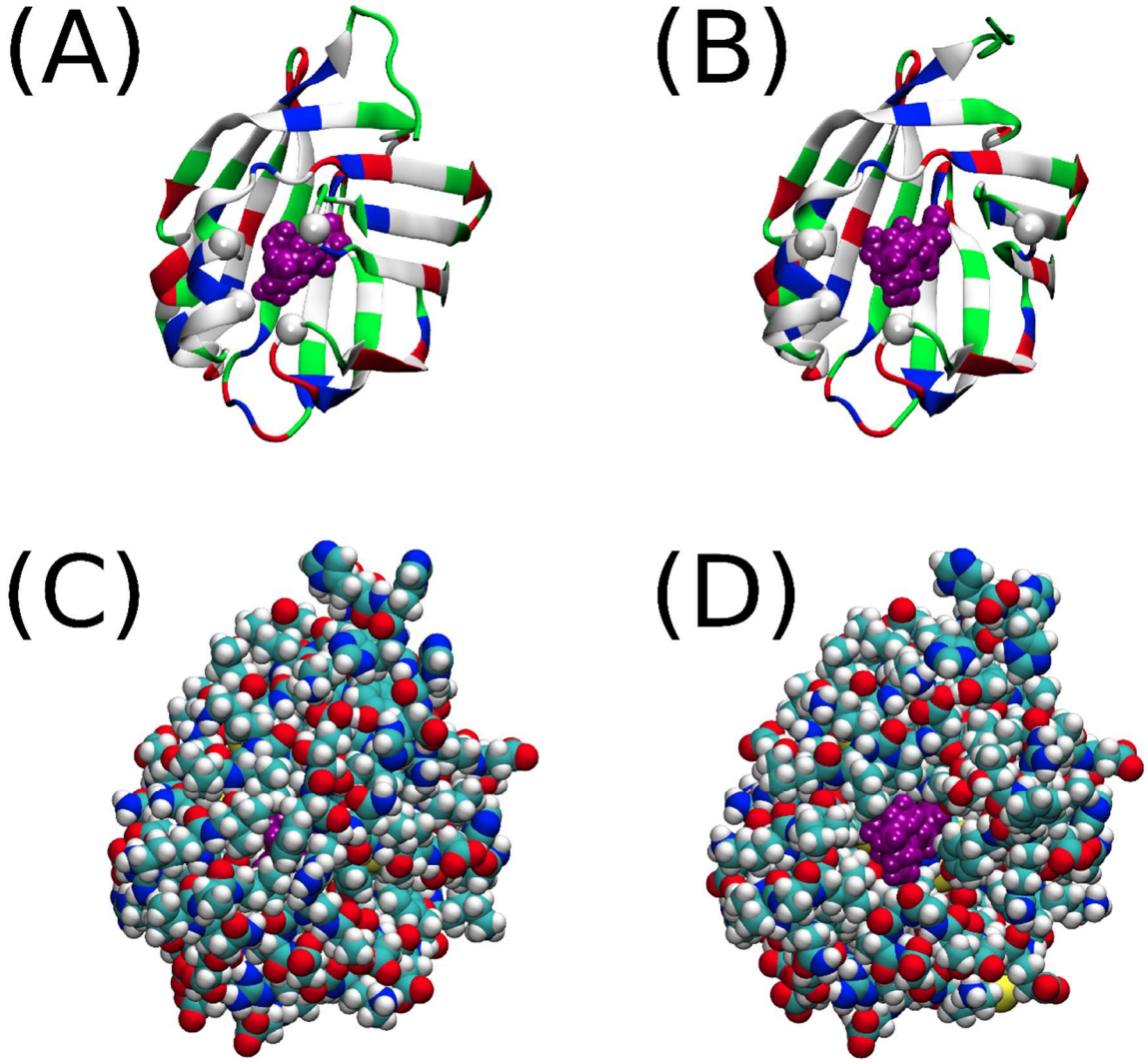

Fig. S1. Complex of retinol inside human cellular retinol-binding protein 1 (RTL-CRBP1). Illustrated in (A) and (C) is the structure of the ligand-protein complex after 50 ns equilibration in physiological saline starting from the X-ray structure (PDB: 5HBS). In (B) and (D), the loop between the beta strands C and D (Residues 54 to 59) is pulled against the alpha helix II (Residues 27 to 35) for 12 Å so that RTL is exposed for extraction from CRBP1. In (A) and (B), CRBP1 is shown in Ribbons colored by residue types (hydrophilic, green; hydrophobic, white; positively charged, blue; negatively charged, red). Four alpha carbons (of Leu 29, Ile 32, Phe 57, Ile 77), shown as large white spheres, are chosen as pulling centers for hSMD simulations. In (C) and (D), CRBP1 is shown as large spheres colored by atom names (H, white; C, cyan; N, blue; O, red; S, yellow). In all panels, RTL is shown in large spheres colored purple. The C6 and C11 of RTL are chosen as the two other pulling centers for hSMD runs.

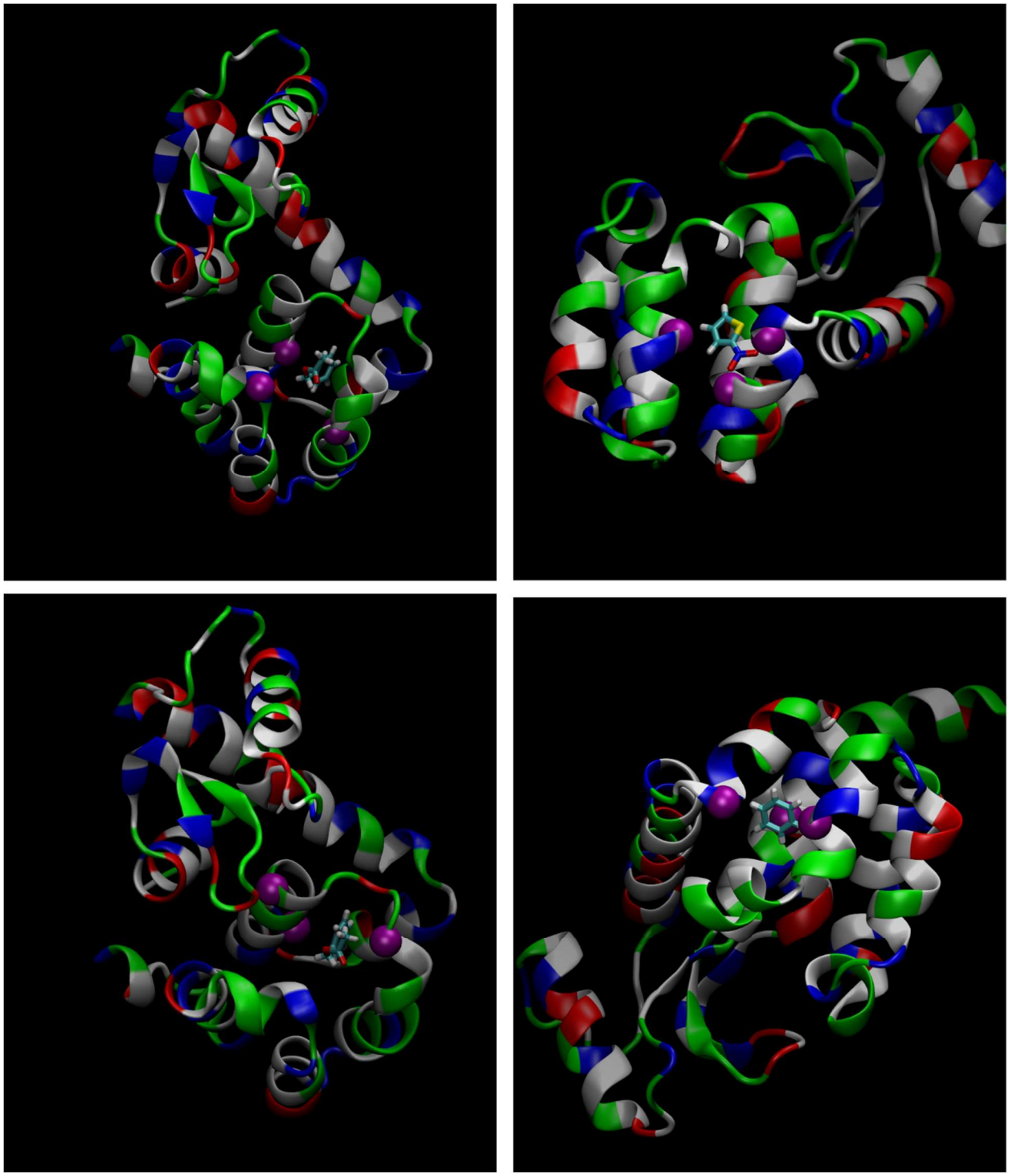

Fig. S2. The T4 lysozyme with the studied ligands. Spheres indicate pulling centers on the protein. Top left, Benzylacetate (J0Z) set I; Bottom left, J0Z set II; Top right, 2-Nitrothiophene (265); Bottom right, Benzene (BNZ).

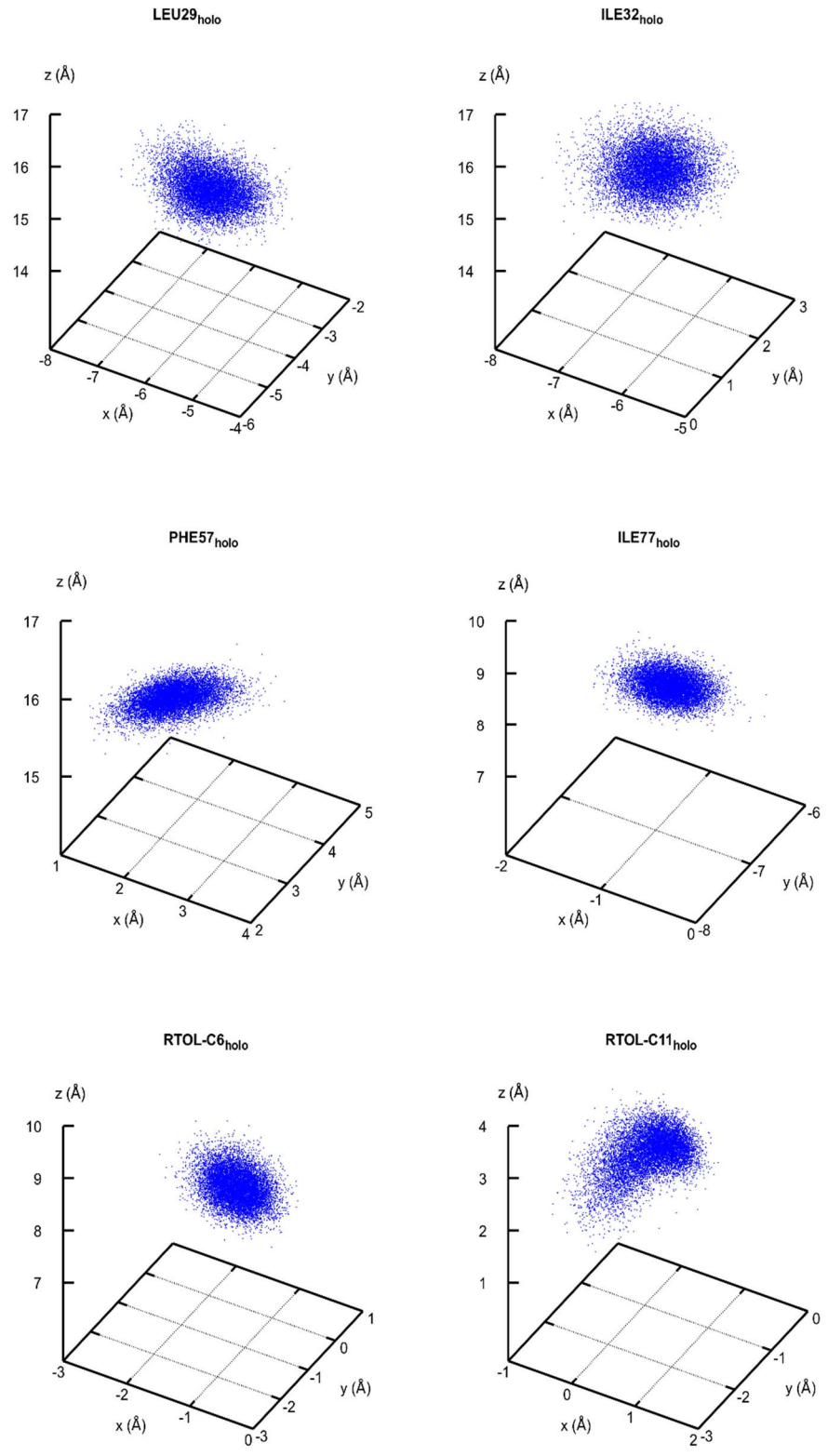

Fig. S3. Fluctuations of pulling centers in the holo state of the RTL-CRBP1 complex.

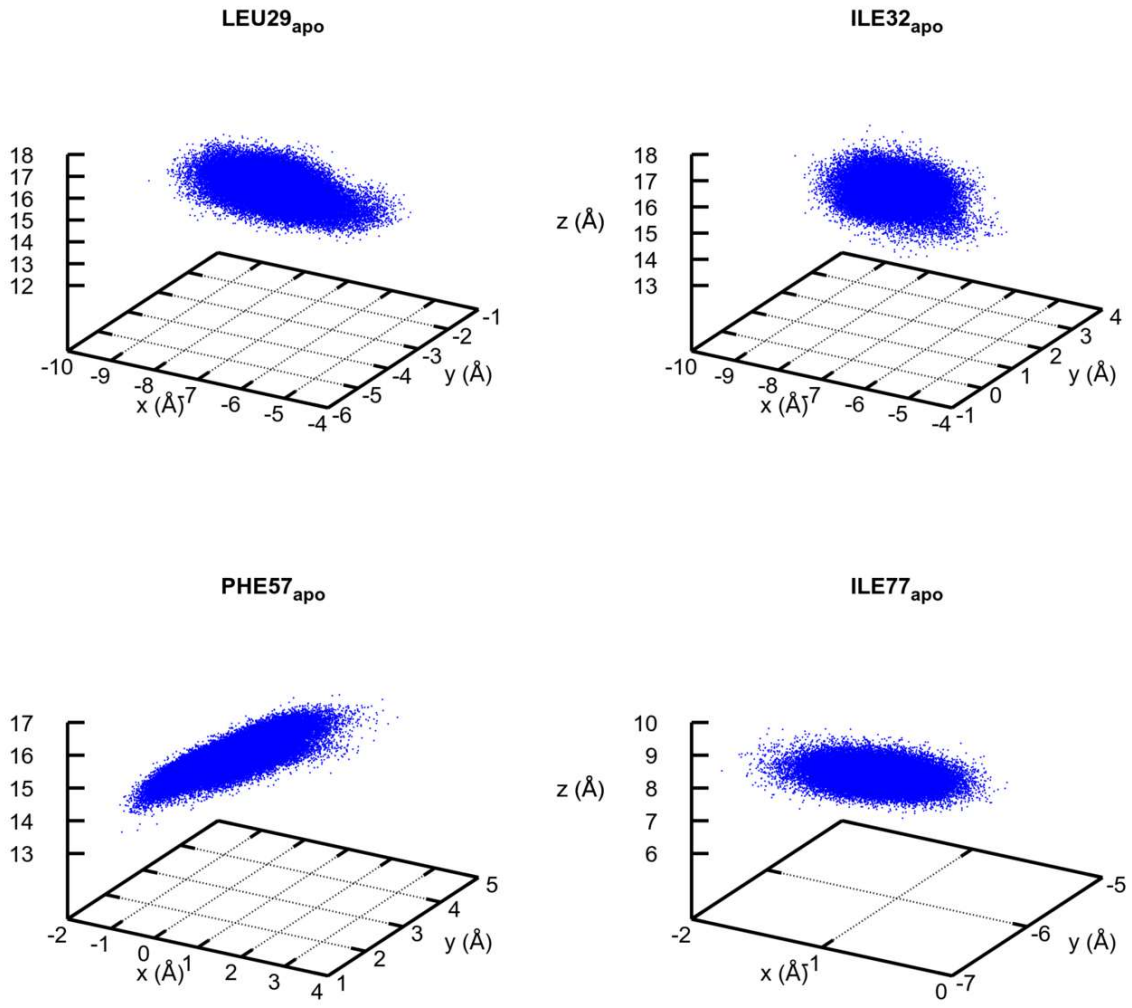

Fig. S4. Fluctuations of pulling centers in the apo state of the RTL-CRBP1 complex.

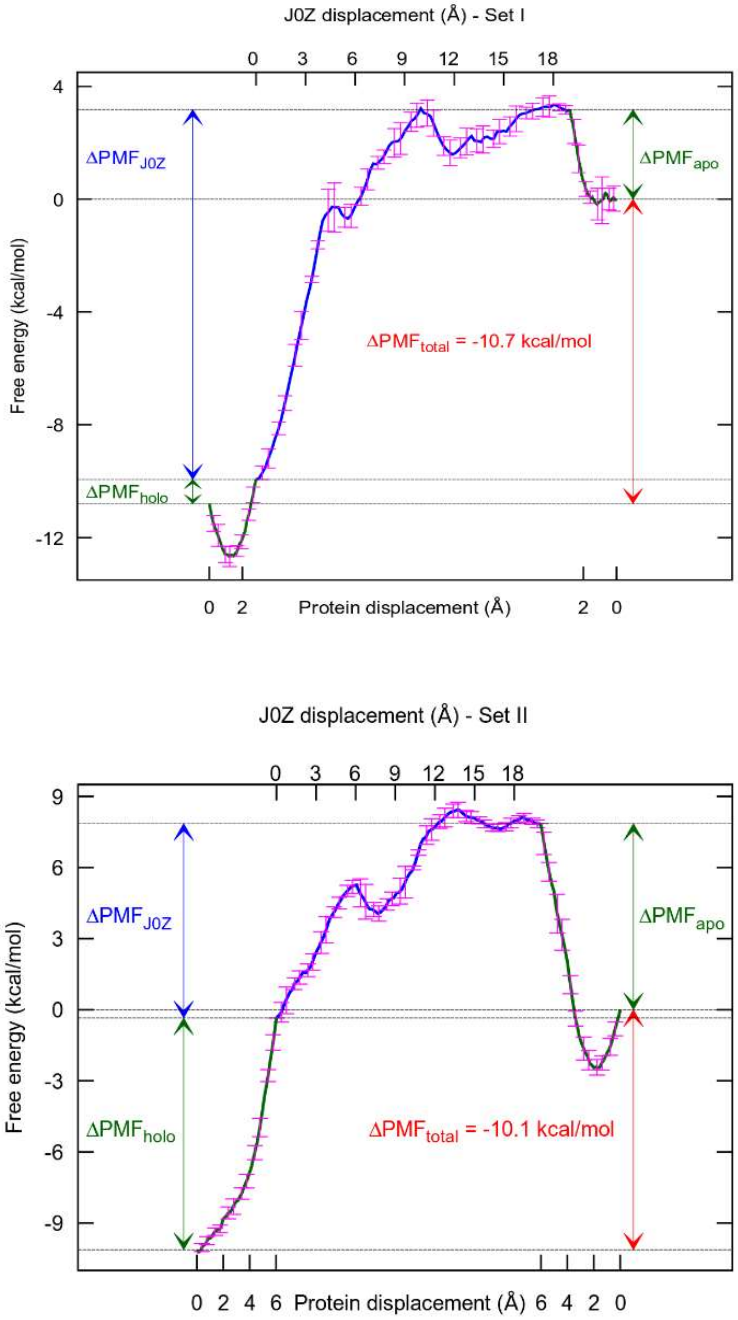

Fig. S5. Two sets of PMF curves along the path of opening up the protein (T4L L99A/M102Q double mutant), pulling out J0Z, and closing (reversely opening) the protein. The pulling centers and the corresponding velocities are in Table I. The largest of the error bars in PMF is taken as the error bar for the binding free energies in Table II. The error bars represent the standard deviations.

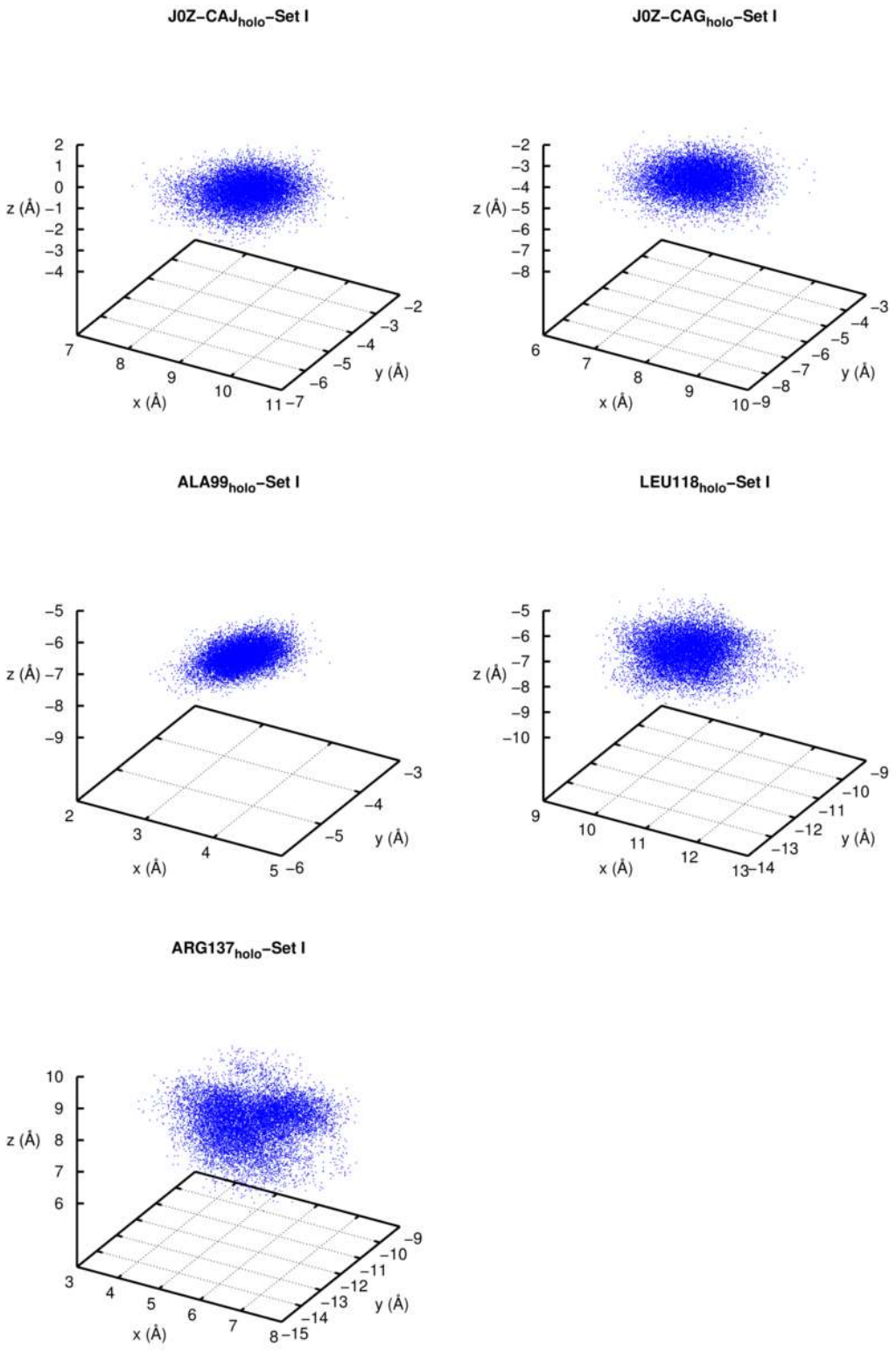

Fig. S6. Fluctuations of pulling centers in the holo state of the J0Z-T4L complex.

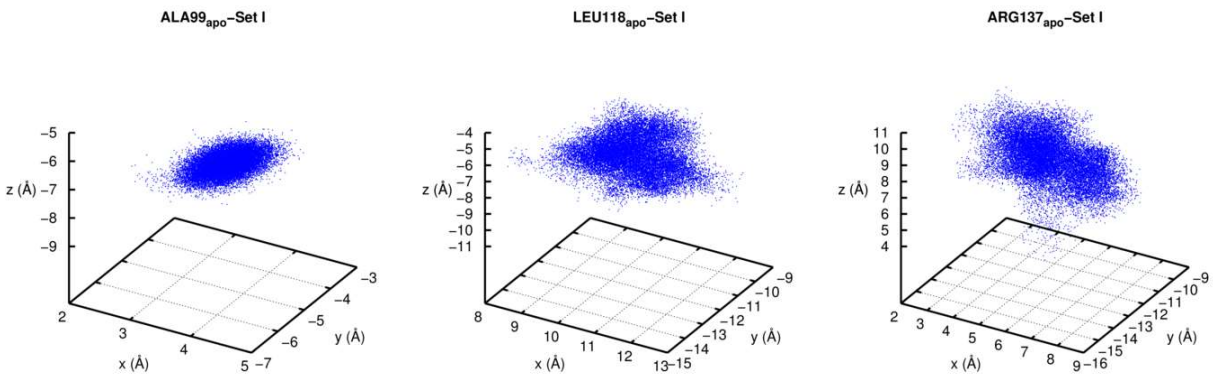

Fig. S7. Fluctuations of pulling centers in the apo state of the J0Z-T4L complex.

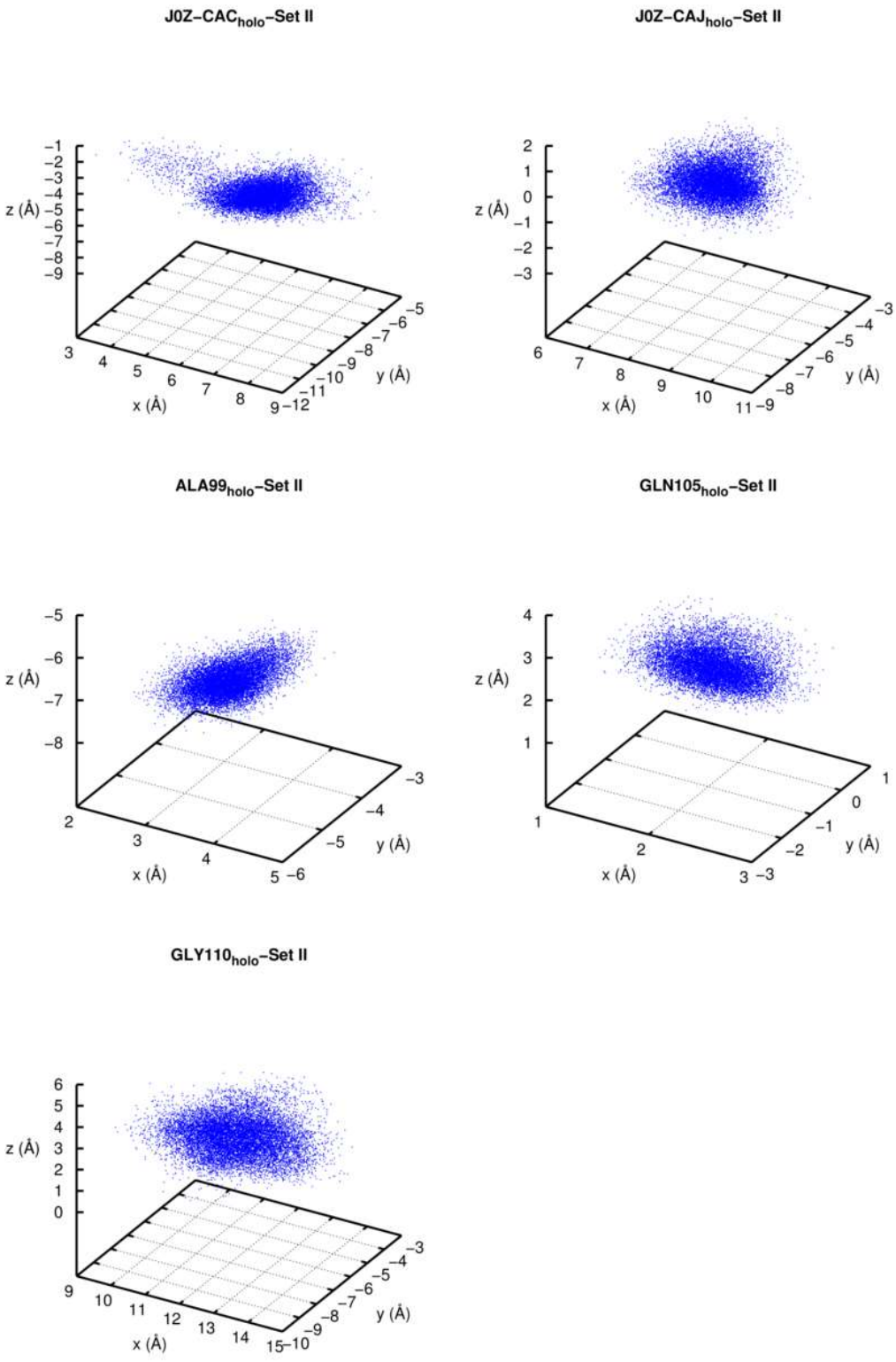

Fig. S8. Fluctuations of pulling centers in the holo state of the J0Z-T4L complex.

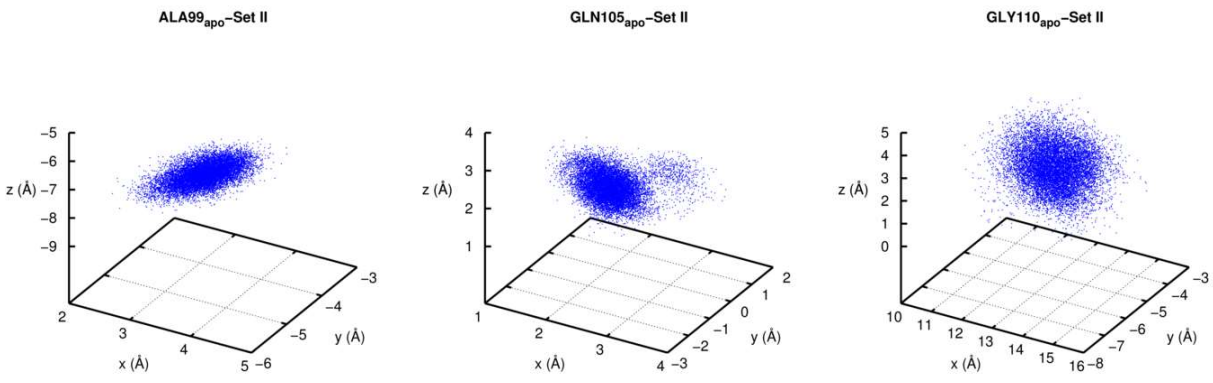

Fig. S9. Fluctuations of pulling centers in the apo state of the J0Z-T4L complex.

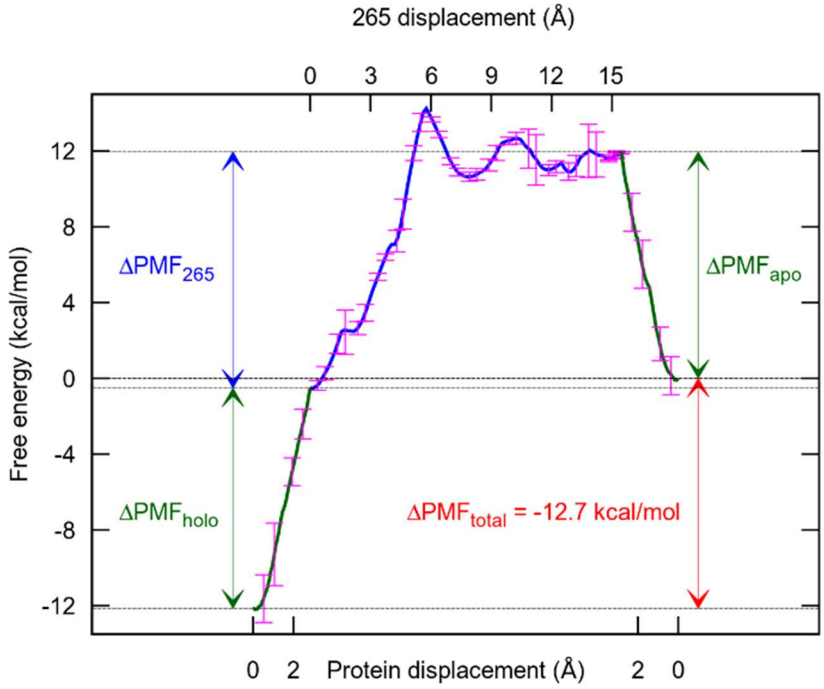

Fig. S10. PMF curves along the path of opening up the protein (T4L L99A/M102Q double mutant), pulling out 265, and closing (reversely opening) the protein. The pulling centers and the corresponding velocities are in Table I. The largest of the error bars in PMF is taken as the error bar for the binding free energies in Table II. The error bars represent the standard deviations.

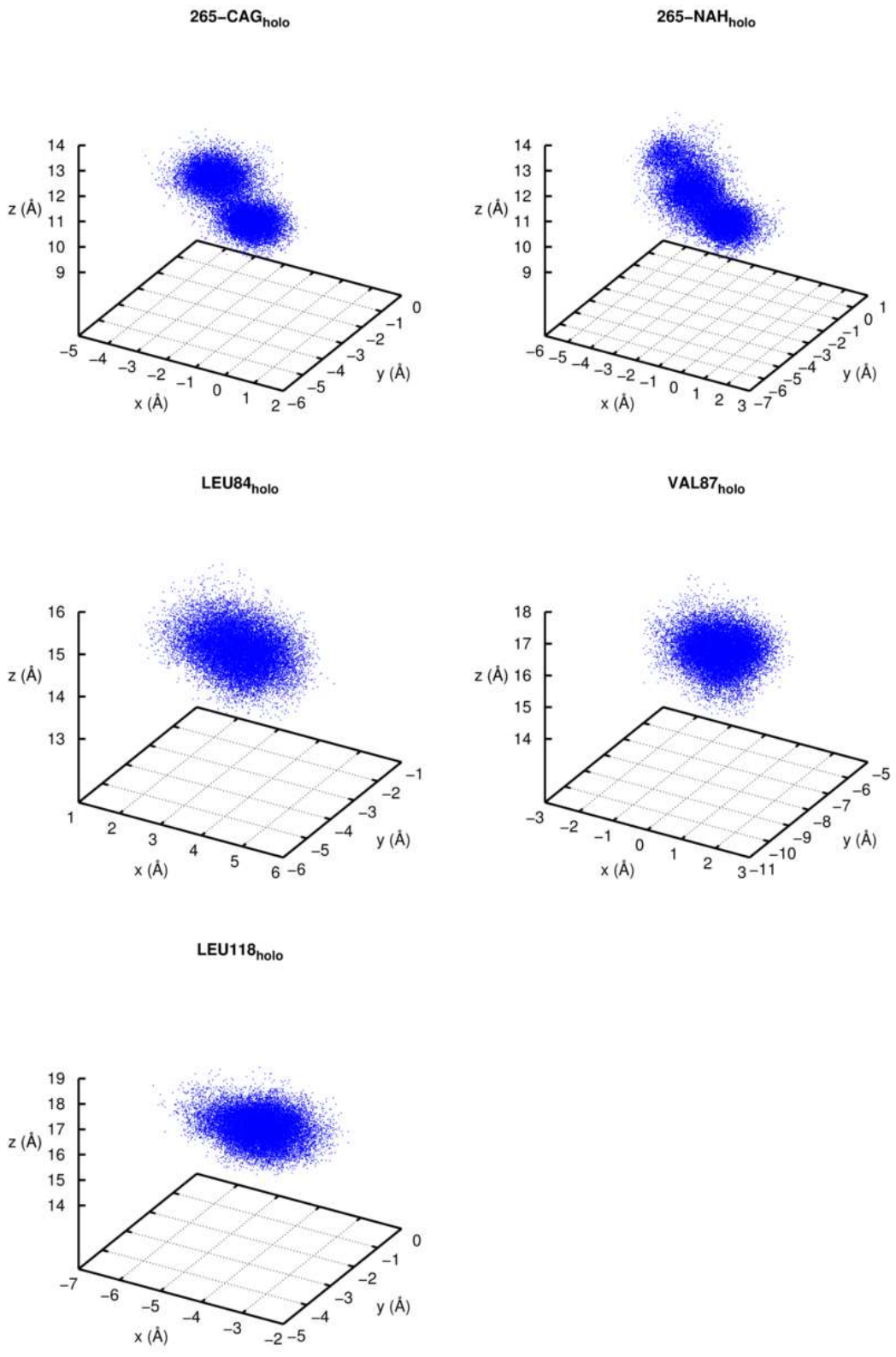

Fig. S11. Fluctuations of pulling centers in the holo state of the 265-T4L complex.

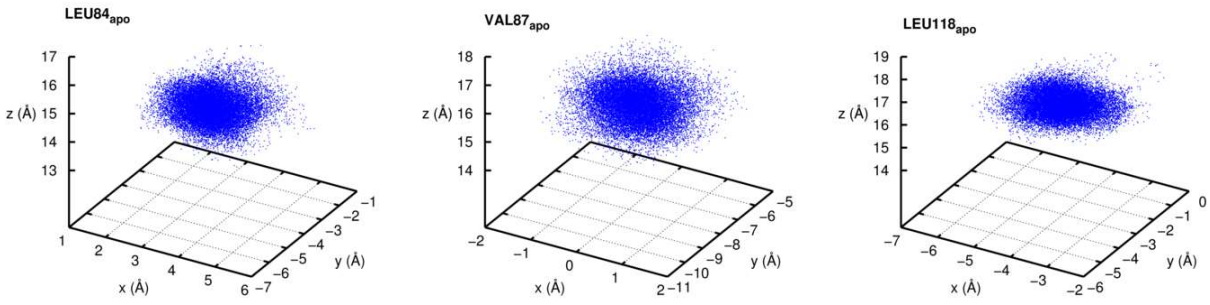

Fig. S12. Fluctuations of pulling centers in the apo state of the 265-T4L complex.

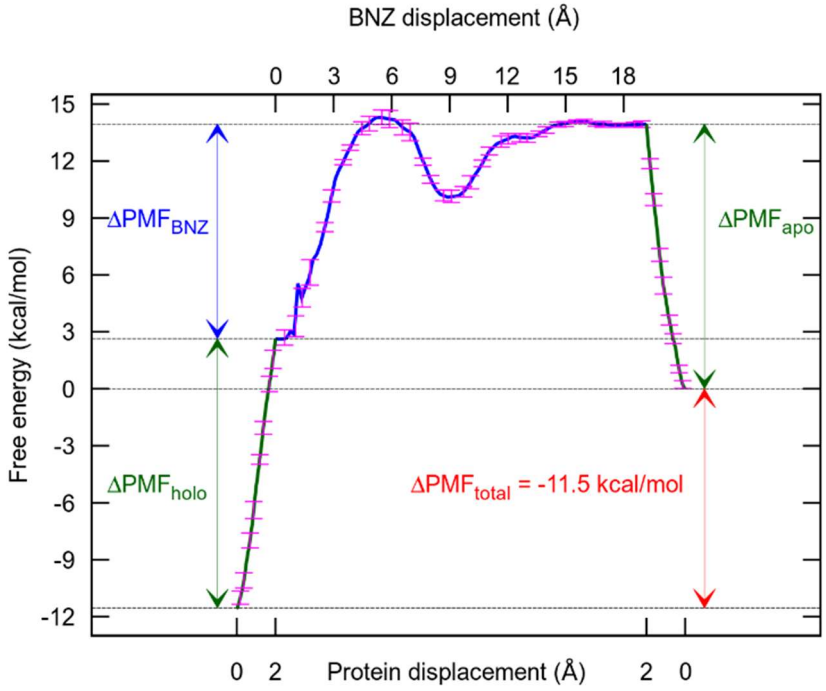

Fig. 13. PMF curves along the path of opening up the protein (T4L L99A/M102H double mutant), pulling out BNZ, and closing (reversely opening) the protein. The pulling centers and the corresponding velocities are in Table I. The largest of the error bars in PMF is taken as the error bar for the binding free energies in Table II. The error bars represent the standard deviations.

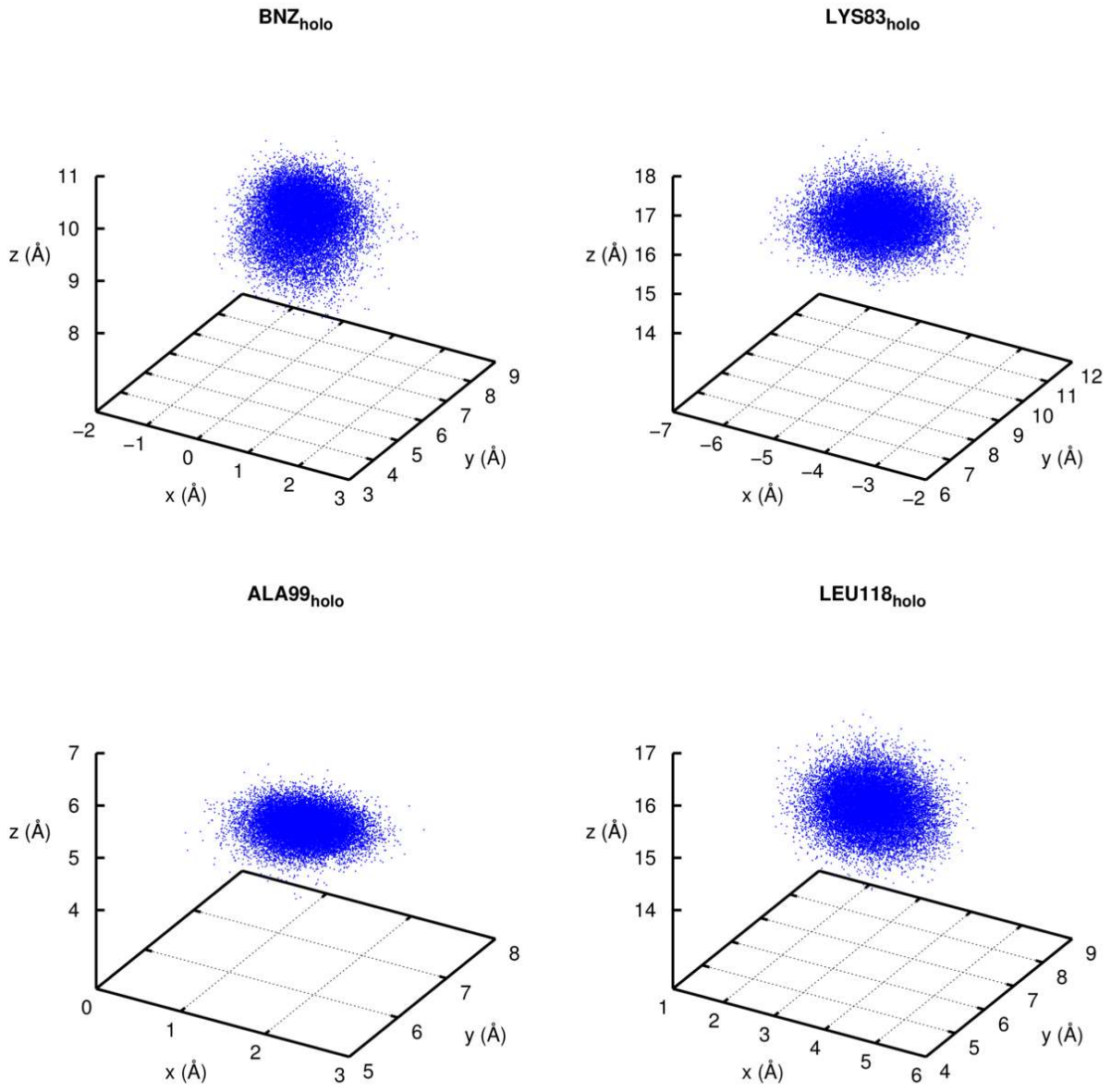

Fig. S14. Fluctuations of pulling centers in the holo state of the BNZ-T4L complex.

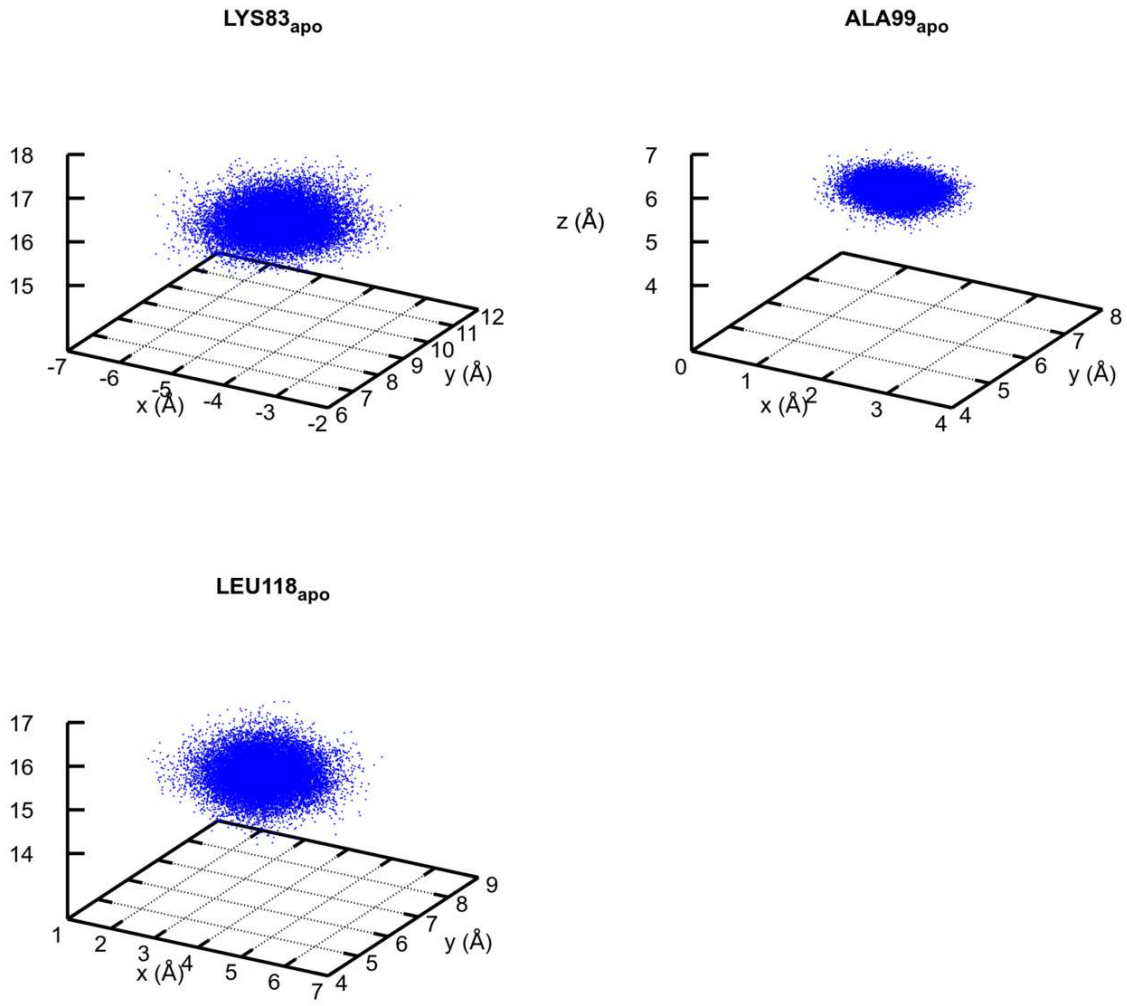

Fig. S15. Fluctuations of pulling centers in the apo state of the BNZ-T4L complex.

REFERENCES

[1] H.-X. Zhou, M.K. Gilson, Theory of Free Energy and Entropy in Noncovalent Binding, Chem. Rev., 109 (2009) 4092-4107.
[2] H.-J. Woo, B. Roux, Calculation of absolute protein–ligand binding free energy from computer simulations, Proc. Natl. Acad. Sci. U. S. A., 102 (2005) 6825-6830.
[3] L.Y. Chen, Hybrid Steered Molecular Dynamics Approach to Computing Absolute Binding Free Energy of Ligand–Protein Complexes: A Brute Force Approach That Is Fast and Accurate, J. Chem. Theory Comput., 11 (2015) 1928-1938.
[4] B. Isralewitz, S. Izrailev, K. Schulten, Binding pathway of retinal to bacterio-opsin: A prediction by molecular dynamics simulations, Biophys. J., 73 (1997) 2972-2979.
[5] C. Velez-Vega, M.K. Gilson, Overcoming Dissipation in the Calculation of Standard Binding Free Energies by Ligand Extraction, J. Comput. Chem., 34 (2013) 2360-2371.
[6] L.Y. Chen, Glycerol modulates water permeation through Escherichia coli aquaglyceroporin GlpF, Biochimica et Biophysica Acta (BBA) - Biomembranes, 1828 (2013) 1786-1793.
[7] L.Y. Chen, D.A. Bastien, H.E. Espejel, Determination of equilibrium free energy from nonequilibrium work measurements, Phys. Chem. Chem. Phys., 12 (2010) 6579-6582.