# SUPPORTING MATERIAL

# Characterization of Protein Flexibility using Small-angle X-ray Scattering and Amplified Collective Motion Simulations

Bin Wen[1, #], Junhui Peng[1, #], Xiaobing Zuo[2], Qingguo Gong[1], and Zhiyong Zhang[1, *]

[1]Hefei National Laboratory for Physical Science at Microscale and School of Life Sciences, University of Science and Technology of China, Hefei, Anhui 230026, People's Republic of China; [2]Advanced Photon Source, Argonne National Laboratory, Chicago, IL 60437

[#]Bin Wen and Junhui Peng contributed equally to this work.

*Corresponding author: Zhiyong Zhang, Tel: +86-551-63600854; Email: zzyzhang@ustc.edu.cn

Running title: Protein Flexibility in Solution

# SUPPLEMENTARY RESULTS AND DISCUSSION

## Convergence of ACM in fitting the SAXS data

### T4L

We have investigated the issue of convergence by running multiple ACM simulations as follows.

*Different starting structures*. We have carried out an ACM simulation starting from a closed structure of T4L. The protein can transit between the closed and the open states back and forth within the 20 ns simulation time (Fig. S2a), like in the ACM simulation starting from the open structure (Fig. 3). EOM yields a structure ensemble containing both closed and open conformations of T4L, with the minimal $\chi$ of 0.008 (Fig. S2b).

*Number of collective modes to be accelerated*. This should be determined based on how many ENM modes are needed to describe the hinge-bending domain motions of T4L. We computed the overlap between the low-frequency ENM modes and the open-close/twist mode, respectively. The first three ENM modes have already shown a good convergence to significantly cover the collective domain motions of T4L (Fig. S3), with an overlap coefficient of 0.89 to the open-close mode and 0.81 to the twist mode (note that a coefficient of 1.0 means complete coverage). Therefore the ACM simulation using the three modes should be better than that using the two modes. On the other hand, there is a technical issue that prevents us from using very few (two or even one) collective modes. In this case, the temperature of one or two degrees of freedom would fluctuate wildly, which may distort the protein structure when the temperature is extremely high (see below the discussion of how to set the high temperature in ACM). For different proteins, the number of collective modes to be accelerated should be system dependent, but we suggest starting with three modes, and adding more if necessary. We have carried out an ACM simulation that coupled the first four collective modes at 800 K, which shows a larger sampling area in the essential subspace (Fig. S4a) than the three-mode ACM simulation does (Fig. 3). However, the EOM ensembles of the two simulations are rather similar (Fig. S4b and Fig. 4c).

*High temperature for ACM coupling*. We performed several ACM simulations, which couple the first three collective modes to different temperatures, respectively. If the temperature is not high enough, the protein cannot cross the energy barrier and reach the closed state, so the ensemble selected by EOM does not fit the SAXS data quite well (data not shown). Figure S5 shows the results of the ACM simulation at 1000 K, which samples a broader region in the essential subspace (Fig. S5a) than the ACM simulation at 800 K does (Fig. 3). The two ACM simulations yield very similar EOM ensembles that contain not only the closed but also the open conformations of T4L (Fig. S5b and Fig. 4c). Generally we have little information on the energy barriers of the protein, so it is not straightforward to determine an optimal temperature for ACM coupling. We usually try a relatively high temperature firstly in order to obtain efficient sampling, but it should be noted that a very high temperature may distort the local structures of the protein since there exists a leakage of

energy between the high-temperature degrees of freedoms to the room-temperature ones. In this case, the temperature should be decreased.

*Different simulation times*. We have extended the 20 ns ACM simulation of T4L to 40 ns. It is found that the 40 ns simulation covers a larger area in the essential subspace (Fig. S6a) than the 20 ns simulation does (Fig. 3), but their EOM ensembles are quite similar (Fig. S6b and Fig. 4c).

## FBP21-WWs

We have finished a series of ACM simulations of FBP21-WWs, as those of T4L. Despite their differences, the EOM ensembles from various trajectories share some similar clusters of structures including both the compact and the extended conformations of the protein (Fig. 6c and Fig. S7). The results again indicate a fairly good convergence of ACM in fitting the SAXS data.

# SUPPLEMENTARY FIGURES

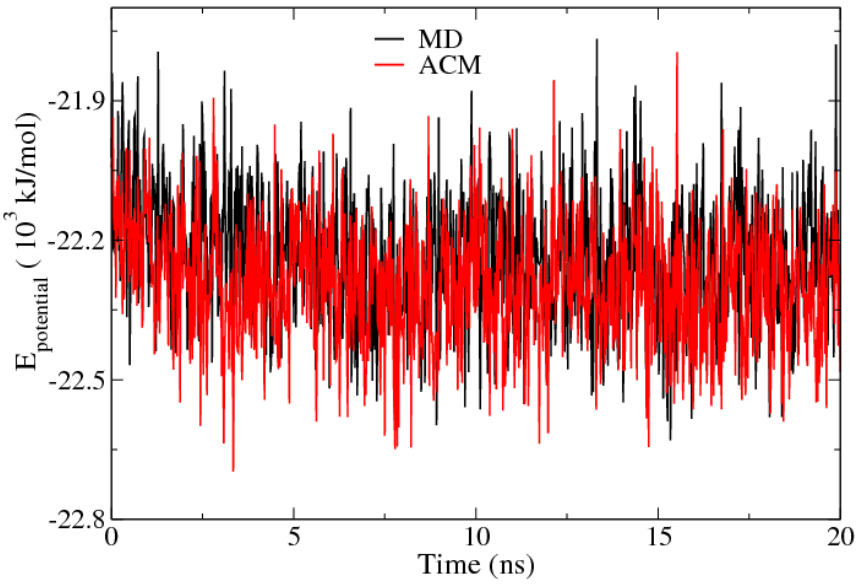


**Figure S1**. Potential energies of the MD (black solid line) and the ACM (red solid line) simulation, respectively. For each trajectory, the explicit water molecules in each frame were removed, and then the solvent contribution was estimated by using an implicit solvent model called the generalized Born surface area (GBSA) model. The calculations were done by using the "-rerun" option of the "mdrun" program in GROMACS-4.5.5 package. In the mdp file, the option "GBSA" was turned on.

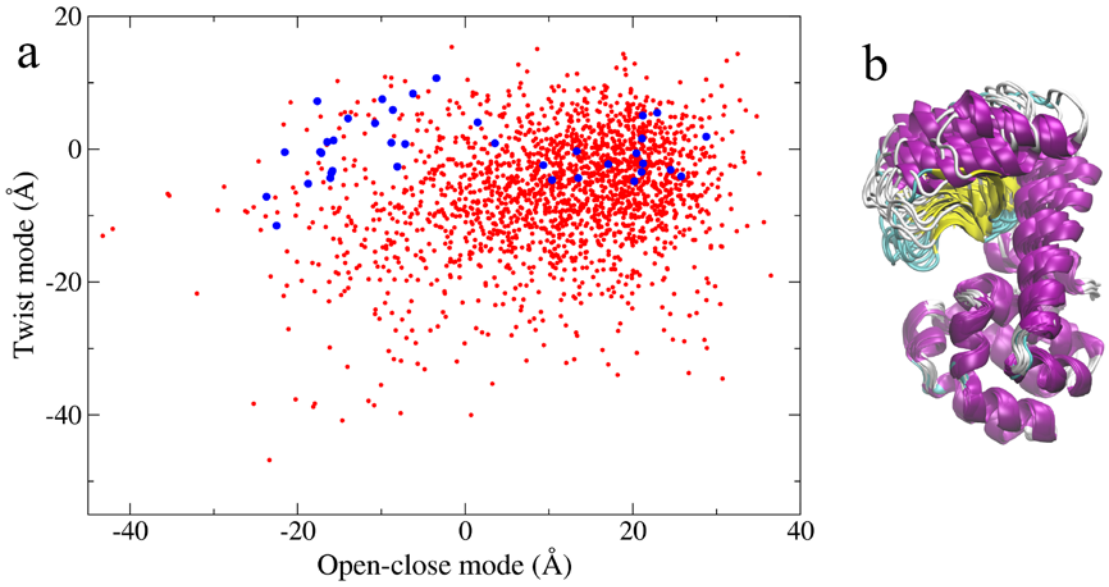


**Figure S2**. The ACM simulation of T4L starting from a closed structure. (a) Conformations in the trajectory are projected onto the 2D essential subspace (colored by red), and the 38 experimental structures of T4L are also show (colored by blue). (b) The structure ensemble selected by EOM with the minimal $\chi$=0.008.

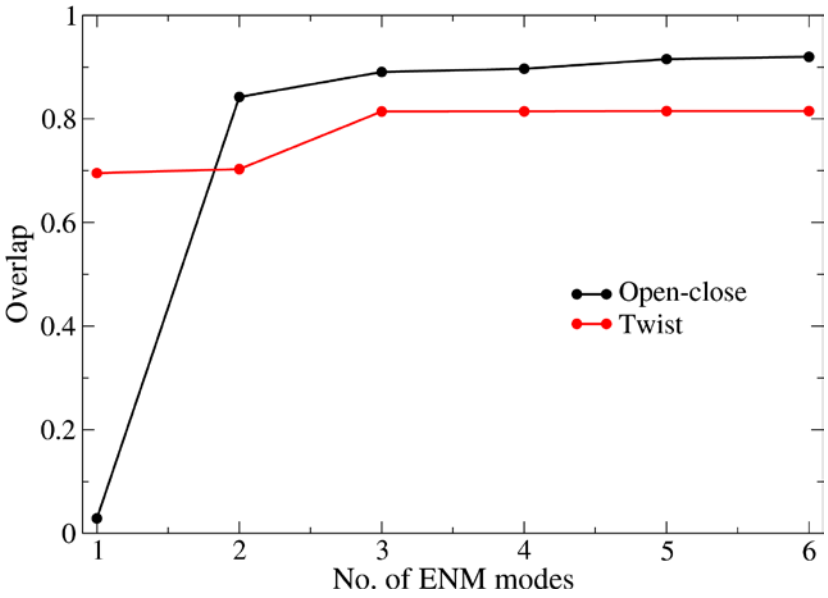


**Figure S3**. Overlap between the ENM modes and the open-close/twist mode of T4L, respectively. For the open/close or the twist mode, we projected it on the subspace formed by the slowest ENM modes (including from one to six modes, respectively), and obtained the overlap values.

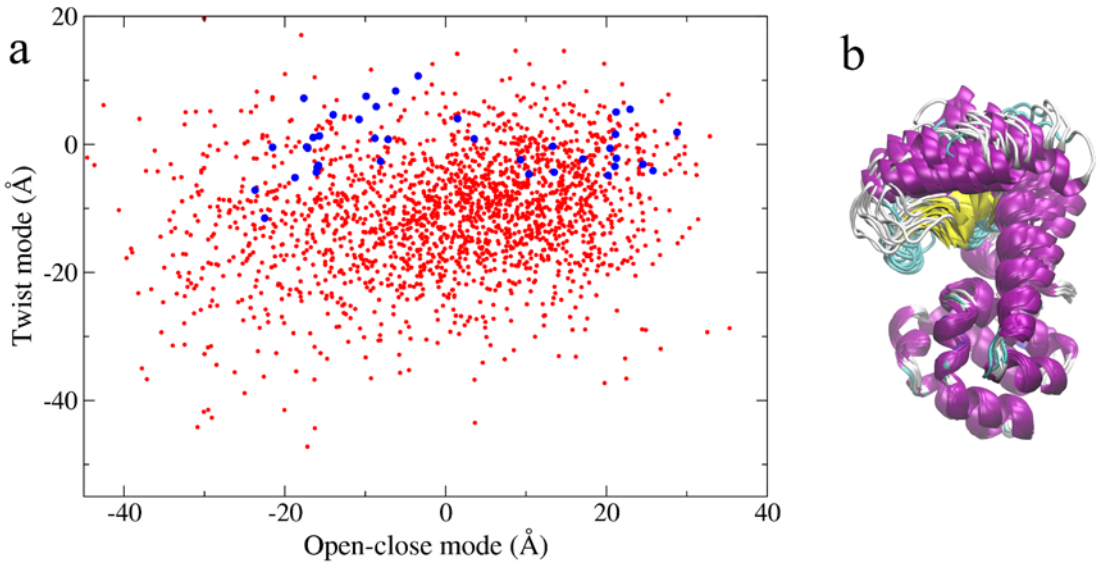


**Figure S4**. The ACM simulation of T4L that couples the first four ENM modes at 800 K. (a) Conformations in the trajectory are projected onto the 2D essential subspace (colored by red), and the 38 experimental structures of T4L are also show (colored by blue). (b) The structure ensemble selected by EOM, with the minimal $\chi$=0.007.

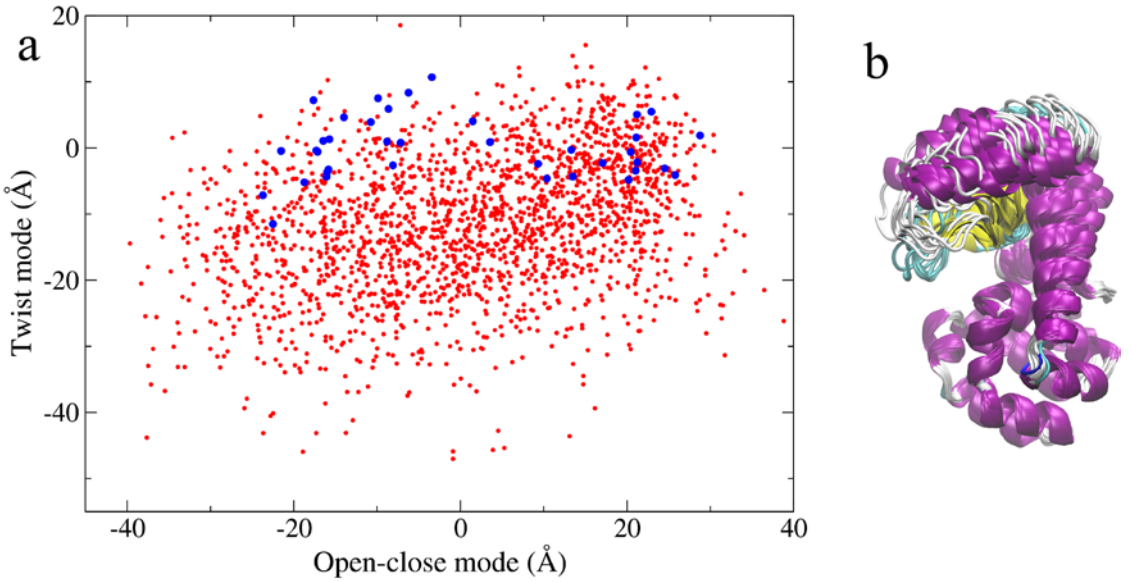


**Figure S5**. The ACM simulation of T4L that couples the first three ENM modes at 1000 K. (a) Conformations in the trajectory are projected onto the 2D essential subspace (colored by red), and the 38 experimental structures of T4L are also show (colored by blue). (b) The structure ensemble selected by EOM, with the minimal $\chi$=0.008.

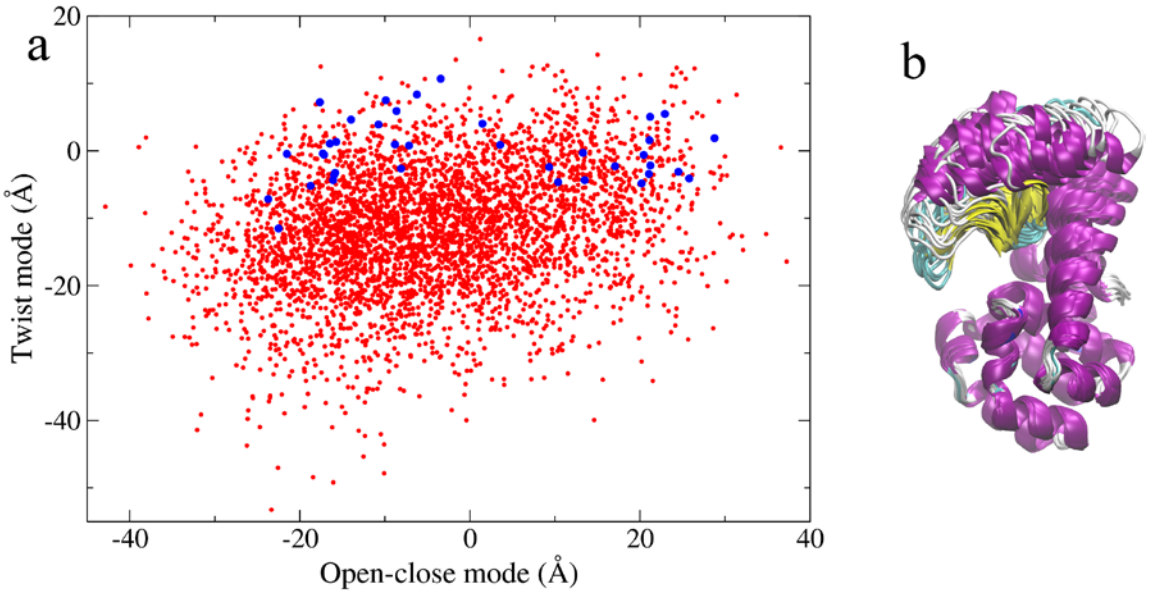


**Figure S6**. The ACM simulation of T4L with a simulation time of 40 ns. (a) Conformations in the trajectory are projected onto the 2D essential subspace, and (b) the structure ensemble selected by EOM, with the minimal $\chi$=0.006.

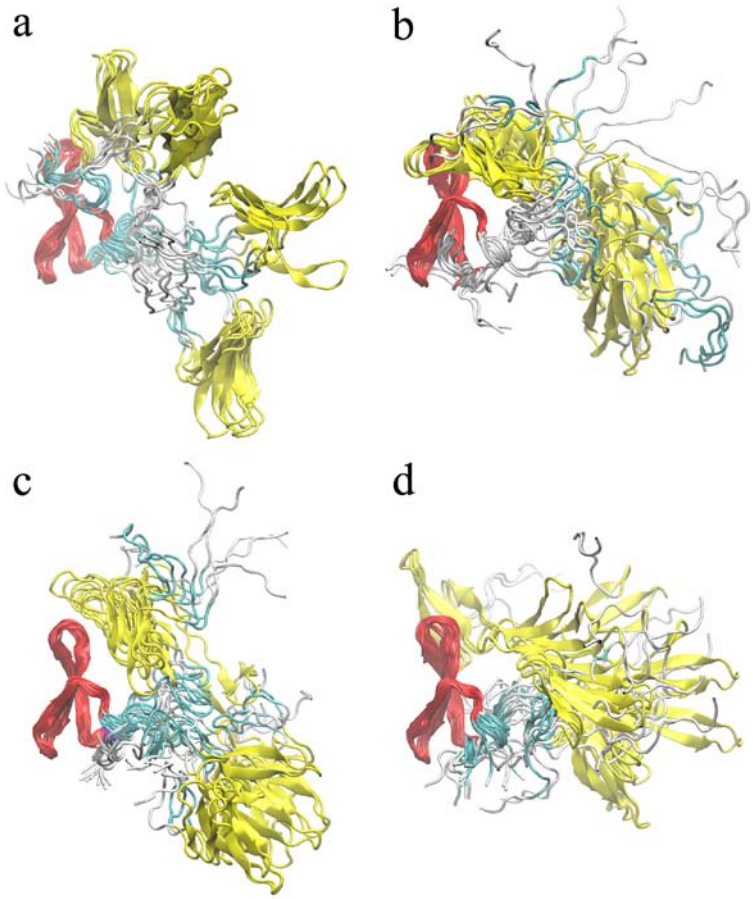


**Figure S7**. EOM ensembles from multiple ACM simulations of FBP21-WWs. (a) Structure ensemble with the minimal $\chi$=0.165 from the ACM simulation starting from an extended structure of the protein. The simulation parameters were the same as those for Figure 6c. (b) Structure ensemble with the minimal $\chi$=0.165 from the ACM simulation that coupled the first four ENM modes at 500K. (c) Structure ensemble with the minimal $\chi$=0.164 from the ACM simulation that coupled the first three ENM modes at 400 K. (d) Structure ensemble with the minimal $\chi$=0.170 from a 40 ns ACM simulation that is an extension of the original 20 ns simulation (Fig. 6c). The structures are superimposed by the WW1 domain (residues 6-32, colored by red), and the WW2 domain (residues 47-73) is colored by yellow.

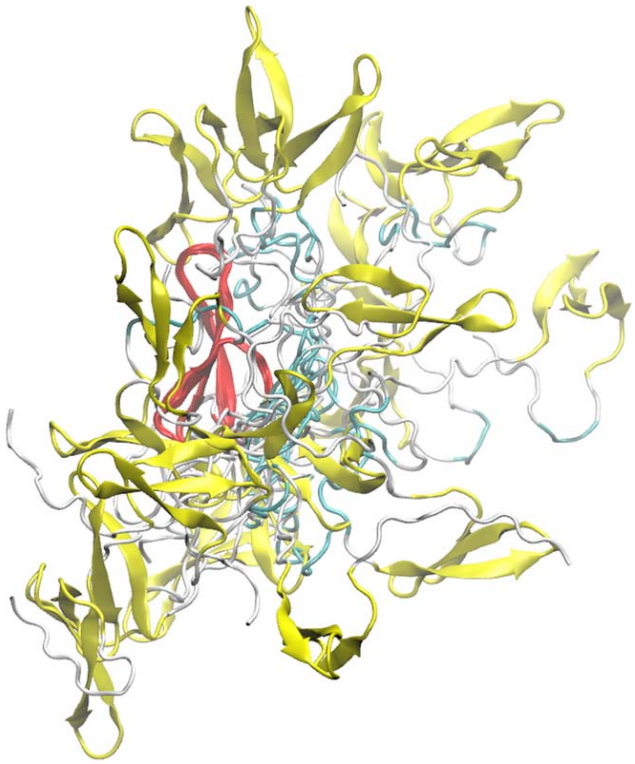

**Figure S8**. EOM ensemble of FBP21-WWs from the structure pool generated by Pre_bunch, with the minimal $\chi$=0.164. The structures are superimposed by the WW1 domain (residues 6-32, colored by red), and the WW2 domain (residues 47-73) is colored by yellow.