Supporting Information

Enhanced Conformational Sampling with an Adaptive Coarse-Grained Elastic Network Model Using Short-Time All-Atom Molecular Dynamics

Ryo Kanada[†*], *Kei Terayama*[‡*], *Atsushi Tokuhisa*[†], *Shigeyuki Matsumoto*[∥] *and Yasushi Okuno*[†∥]

[†]RIKEN Center for Computational Science, Kobe 650-0047, Japan

[‡]Graduate School of Medical Life Science, Yokohama City University, Yokohama 230-0045, Japan

[∥]Graduate School of Medicine, Kyoto University, Kyoto 606-8507, Japan

[*]Corresponding Authors

Supporting Information: Methods

S1 Q-score: Contact Fraction of Native Contact Pairs

The order parameter Q-score is the contact fraction of the formed native contact pairs in a given conformation. The native contact pairs are defined based on the reference PDB structure with the distance threshold $d_T = 6.5$ Å as mentioned in **Supporting Information S3**. For a given protein conformation, we consider that the native contact between the i-th and j-th amino acids is formed if the distance r_{ij} is < 1.2 $r_{ij}^{(0)}$, where $r_{ij}^{(0)}$ is the distance for the ij pair at the reference structure in protein data bank (PDB). Therefore, the Q-score indicates the degree to which the domain of the native structure is stably maintained at a given structure, i.e., the closer the Q-score to 1, the more stable and similar to the reference structure.

In this study, Q-score (apo) and Q-score (holo) represent the contact fraction of the formed native contact pairs based on the apo and holo structures, respectively.

S2 All-atom Molecular Dynamics (AA-MD) Simulation

To prepare the dynamic cross-correlation coefficient map (DCCM)[1], starting from the crystal structure in the apo-state (pdb-id:4AKE for ADK and pdb-id:1WDN for GBP), energy minimization and All-atom Molecular Dynamics (AA-MD) simulations in an explicit solvent were conducted using GROMACS version 4.6.5[2] using the AMBER ff99SB-ILDN force field[3] as described below.

An octahedron simulation box was constructed with a margin of approximately 8–10 Å to the boundary of the box. The target molecule was solvated with approximately 150 mM of NaCl solution, composed of TIP3P water molecules[4] and sodium and chloride ions, using the GROMCS genion module, which can neutralize the net charge of the simulation system. Following energy minimization using the steepest descent algorithm, the system was equilibrated for 100 ps under a constant volume and temperature (NVT) ensemble at T = 298 K, followed by an MD run for 100 ps under constant pressure and temperature (NPT) conditions at T = 298 K and P = 1 atm with positional restraints applied on protein heavy atoms using a Nose-Hoover thermostat[5] with a time constant of 0.3 ps and a Berendsen barostat[6] with a time constant of 1.0 ps. Following equilibrium MD, production runs were conducted under the NPT ensemble at T = 298 K and P = 1 atm without positional restraints using the same thermostat and barostat as for equilibrium MD. During AA-MD simulation, under the periodic boundary condition, the van der Waals forces were switched smoothly to zero over a range of 8–10 Å, and electrostatic interactions were calculated using the particle mesh Ewald method[7] with a cutoff length of 11 Å. The LINCS algorithm[8] was applied to constrain all bond lengths with a simulation time step of 2 fs. To evaluate the DCCM in step 1 of the adaptive coarse-grained Elastic Network Model (ENM), we conducted five independent production runs of at most 50 ns with different initial velocity conditions.

For qualitative analysis of the sampled structures, as shown in "Sampling Performance of New Adaptive CG-ENM and Comparison with Conventional CG-ENM and AA-MD" in the Results and Discussion, a productive run of the AA-MD simulation up to 1 μs was also conducted under NPT conditions (T = 300 K and P = 1 atm). The five productive run trajectories of an order of at most 50 ns, with coordinate frames taken every 2 ps, were used for calculating the DCCM ("Evaluation of DCCM Based on Short-Time AA-MD Simulations" in Results and Discussion. Concretely for DCCM in Figure 2C and 2D, the AA-MD trajectories of 50 ns × 5 and 1.5 ns × 5 were utilized, respectively). To qualitatively analyze the structural ensemble ("Sampling Performance of New Adaptive CG-ENM and Comparison with Conventional CG-ENM and AA-MD"), 5000 structures taken every 0.2 ns from one productive run trajectory of 1 μs were used.Detailed information related to the AA-MD procedure for integrin in inactive state is mentioned in **Supporting Information S7**.

S3 Conventional coarse-grained (CG)-ENM

The force field for conventional CG-ENM (Trion-type CG-ENM) is expressed by the following equation:

$$E = \sum_{i<j}^{nat\ contact} K_{ij}\left(r_{ij} - r_{ij}^{0}\right)^{2} \qquad (\mathrm{S3.1})$$

where K_{ij} (kcal/mol/ Å ^{2}) and r_{ij}(Å) are the spring constant and the distance between the i-th and j-th residue pair, respectively, and r_{ij}^{0} is the distance at the reference structure (apo-structure of ADK and GBP, and inactive-structure of integrin) for the corresponding pair. The summation of the energy term is only performed over the native contact pairs (i-j), which are defined as follows: when one of the heavy atoms in the i-th amino acid is within a distance threshold, d_T = 6.5 Å, from any non-hydrogen atom in the j-th amino acid in the reference structure, the i-j pair is regarded as a native contact pair. In this study, by applying an enhanced sampling method, temperature replica exchange MD (TREMD), we investigated two types of conventional CG-ENMs with the spring constant K_{ij} = 10 (kcal/mol/Å ^{2}), which is the default in CafeMol software, and the 10-fold weaker spring $K_{ij} = 1$ (kcal/mol/Å^2).

S4 CG-MD Simulation with Under-damped Langevin Dynamics

In this study, the time evolution of the CG system is expressed by the following under-damped Langevin dynamics equation:

$$\mathrm{m}_i \frac{d^2 \boldsymbol{r_i}}{dt} = \boldsymbol{f_i} - \mathrm{m}_i \boldsymbol{\gamma_i} \frac{d r_i}{dt} + \mathrm{m}_i \boldsymbol{\xi_i}(t)\,, \tag{S4. 1}$$

where m_i and γ_i are mass and friction coefficients for the i-th residue (coarse-grained particle), $\boldsymbol{f_i}$ is the force derived from the total energy function E as $\boldsymbol{f_i} = -\frac{dE}{dr_i}$, and $\boldsymbol{\xi_i}(t)$ is the white Gaussian noise, which is responsible for the thermal fluctuation of the system through solvent effects. For the mass and friction coefficient, we adopted the default values in CafeMol 3.2: each particle has a residue-type dependent mass m_i and a constant friction coefficient $\gamma_i = 0.8435$. The white Gaussian noise $\boldsymbol{\xi_i}(t)$ satisfies the following equation, termed as the fluctuation-dissipation theorem:

$$\langle \boldsymbol{\xi_i}(t) \rangle = \boldsymbol{0}\,,\ \langle \boldsymbol{\xi_i}(t) \boldsymbol{\xi_j}(t') \rangle = \frac{\boldsymbol{2k_B T}}{\boldsymbol{m_i}} \boldsymbol{\delta}(\mathbf{t} - \mathbf{t}') \boldsymbol{\delta_{ij}}, \tag{S4. 2}$$

where the bracket represents the ensemble average, $\boldsymbol{k_B}$ is the Boltzmann constant, and T is the temperature.

For parameter searching by Bayesian optimization (BO) (or exhaustive search) in step 2 of adaptive CG-ENM, starting from the reference structure (of ADK and GBP in apo-state, and of integrin in inactive-state), we conducted an under-damped Langevin dynamics simulation at T = 300 K of up to 10^7 steps with a time-step dt = 0.2 for each parameter set. To evaluate the BO target function F_{BO}, 5000 structures taken every 2000 steps from one trajectory (of 10^7 steps) were used.

After parameter searching in step 2, with a suitable parameter set tuned by BO, a productive run of adaptive CG-ENM was conducted for 10^7 steps for ADK and GBP and 5×10^7 steps for integrin by applying an under-damped Langevin dynamics simulation at T = 300 K and dt = 0.2 with a different random seed for white Gaussian noise starting from initial structure (in apo-state for ADK and GBP, and in inactive-state for integrin). To analyze the sampled structure ensemble and time evolution of the Root Mean Square Deviation (RMSD), Q-score, radius of gyration (Rg), and other measurements, 5000 structures taken every 2000 steps from one trajectory of 10^7 steps were used for

ADK and GBP, whereas 5000 structures taken every 10000 steps from one trajectory of 5×10^7 steps were used for integrin

S5 Temperature Replica Exchange MD (TREMD) of Conventional CG-ENM

Temperature replica exchange molecular dynamics (TREMD) is a representative enhanced sampling method used to sample various structures broadly. Beginning from the apo-structure for ADK and GBP and from the inactive structure for integrin, we conducted TREMD simulation of conventional CG-ENM for 256 replicas distributed exponentially in the range of 300–1000 K by applying under-damped Langevin dynamics for 10^8 steps with a time step dt = 0.2. We confirmed that during TREMD simulation, temperature exchange occurs with sufficient frequency as shown in **Figure S22**, which illustrates the replica-ID itinerancy at 300 K. To compare the sampled structural ensemble of conventional CG-ENM by TREMD with those of AA-MD and adaptive CG-ENM, 5000 structures taken every 2 × 10^4 steps from a replica trajectory (of 10^8 steps) at 300 K were used.

S6 Comparison with Two Enhanced Sampling Methodologies

We also compared adaptive CG-ENM with two other sampling methodologies that can realize modeling of holo-like structures from apo conformation[9,10].

In Seelinger's procedure[9], biased tCONCOORD[11,12] sampling that takes the apo conformation (including atomistic information such as bond, angle, and dihedral) and the radius of gyration of target (holo) structure as constraints, is followed by the refinement procedure including energy minimization and AA-MD simulation. As a result, their methodology successfully generated the structure models within 1.6 Å backbone RMSD to the target (holo-structure). This RMSD value 1.6 Å is much smaller than RMSD (vs holo) of the structure S1 sampled by adaptive CG-ENM: 3.9 Å for ADK and 4.1 Å for GBP (as seen in Figure 5 and 6). However, our sampling procedure with adaptive CG-ENM does not require any information regarding the target (holo) structure, including the radius of gyration Rg. Therefore, based on only the apo-structure, we compared the structural ensemble sampled by tCONCOORD, which is one of a key element techniques in their broad structural sampling method[9], without using the Rg of the target (holo) as constraints with ensemble sampled by our adaptive CG-ENM. Figures S19 and S20 show the probability distribution of Rg and Cα pair-distance between representative residues 40-149 for ADK and 50-118 for GBP based on 5000 structures sampled by two methods. As shown in panel A of these figures for ADK and GBP, the structural ensemble by our adaptive CG-ENM is significantly broader than the one sampled by tCONCOORD without the Rg of holo as a constraint. Furthermore, model S1, of which RMSD vs holo is significantly smaller among the ensemble structures sampled by our adaptive CG-ENM was compared with model T1, whose RMSD to the target structure is the smallest among the ensemble structures sampled by tCONCOORD, which does not use Rg of holo as a constraint. The Cα pair-distance between representative residues in model-S1 (20.0 Å for ADK and 9.6 Å for GBP) sampled using our method is significantly closer to the corresponding distance in the target-holo (20.3 Å for ADK and 7.5 Å for GBP) than that in model-T1 (33.8 Å for ADK and 14.2 Å for GBP) sampled by tCONCOORD, as shown in panel B in Figures S19 and S20.

Dokainish[10], by applying AA-MD with gREST_SSCR, which is a type of enhanced conformational sampling algorithms, to ribose binding protein (RBP), succeeded in exploring large domain motion such as the open-closed conformational change. By utilizing the advantages of enhanced sampling with AA-MD, they also succeeded in determining important atomistic interactions through hydrogen bond analysis, which, in principle, could not be performed by our adaptive CG-ENM. Their simulation target RBP[13] differs from our targets ADK and GBP; hence, it is difficult to make a direct comparison between the structural ensembles sampled by their gREST_SSCR and our adaptive CG-ENM. However, there are two similarities between RBP and GBP: i) two globular proteins are mainly composed of two domains and ii) the conformational transitions of two proteins between the apo-holo state are mainly caused by hinge motion. Therefore, it is assumed that the diversity of the sampled structural ensemble compared using gREST_SSCR (at T=300.0 K) based on RBP apo-state and adaptive CG-ENM based on GBP apo-state would provide some supporting information related to sampling performance. We selected the relative distribution width of the radius of gyration ($^{\Delta Rg}/_{Rg_{apo}}$) based on the sampled structure ensemble, defined by the difference (ΔRg) between the maximum and minimum values of Rg in the sampling structures and the ratio of Rg in the apo structure, as the measurement for the diversity of sampled ensemble. As a result of estimation, we found that $^{\Delta Rg}/_{Rg_{apo}}$ estimated using our adaptive CG-ENM: 0.24 is wider than that estimated by their gREST_SSCR: 0.15.

S7 Detailed Information related to AA-MD Procedure for Integrin αV in Inactive State

To evaluate DCCM for integrin, beginning from the crystal structure in the inactive state of integrin αV (pdb-id: 1JV2), energy minimization and AA-MD simulations in an explicit solvent were conducted using GROMACS version 4.6.5[2] with the AMBER ff99SB-ILDN force field[3] through the same procedure for ADK and GBP mentioned in Supporting Information S2. For DCCM in Figure 7B, the AA-MD trajectories of 1.5 ns × 5 were utilized（five short AA-MD trajectories were concatenated to provide a single C_{ij} value for each residue pair）. To qualitatively analyze the structural ensemble in the section "Application of the New Adaptive CG-ENM to Larger Protein System Integrin αV" of Result and Discussion), 5000 structures taken every 10 ps from one AA-MD productive run trajectory of 50 ns were used.

Supporting Information Figures:

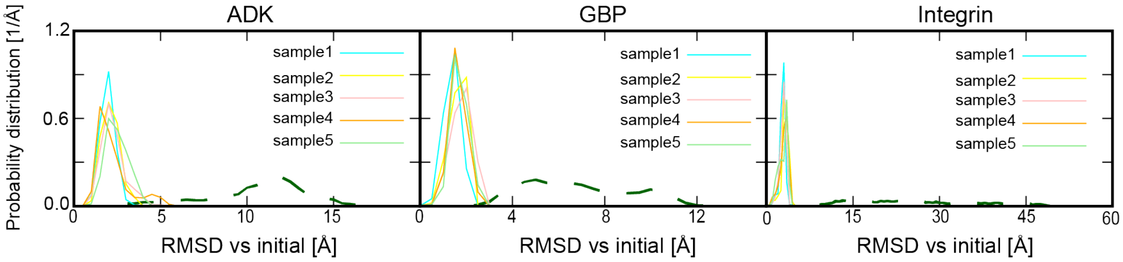

Figure S1. Probability distributions of RMSD vs initial structure for five short-time AA-MD trajectories with different initial velocities for samples 1–5. The probability distributions for adaptive CG-ENM in ADK, GBP, and Integrin are calculated based on structure ensemble sampled by production-run with a suitable parameter set (*Ks*, *Kw*, *Cs*, *Cw*) = (8.0, 7.0, 0.8, 0.6), (10.0, 8.0, 0.8, 0.6), and (7.0, 1.0, 1.0, 0.8) explored by BO. The probability distributions for five short-time AA-MD are shown with colored solid lines (cyan, yellow, pink, orange, light green). The dark-green dashed line shows the probability distribution for adaptive CG-ENM.

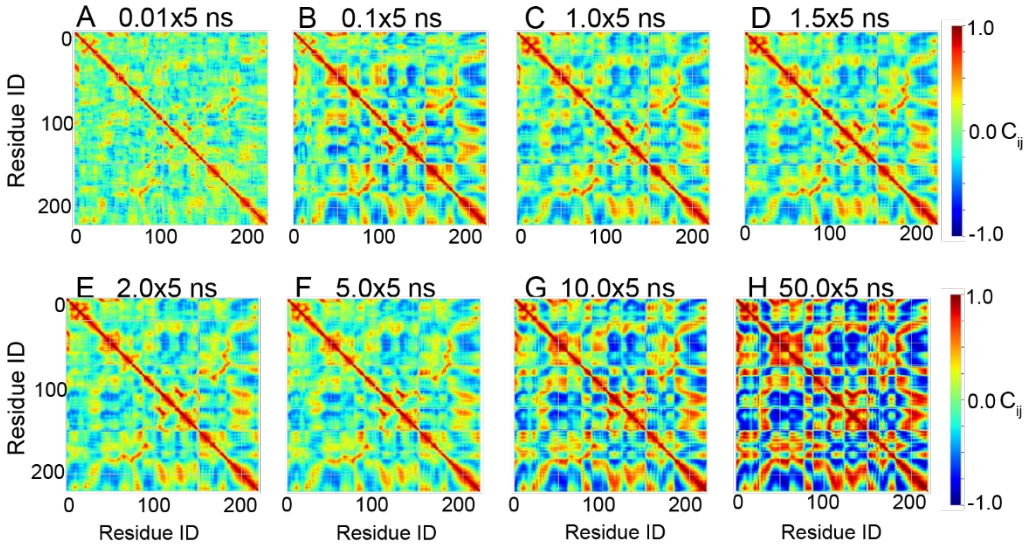

Figure S2. DCCM for GBP based on various time lengths of AA-MD trajectories in the plane of residue-id pairs. Each panel from (A) to (H) corresponds to DCCM by using AA-MD trajectories for 0.01 ns × 5, 0.5 ns × 5, 0.1 ns × 5, 1.0 ns × 5, 1.5 ns × 5, 2 ns × 5, 5 ns × 5, 10 ns × 5, and 50 ns × 5, respectively.

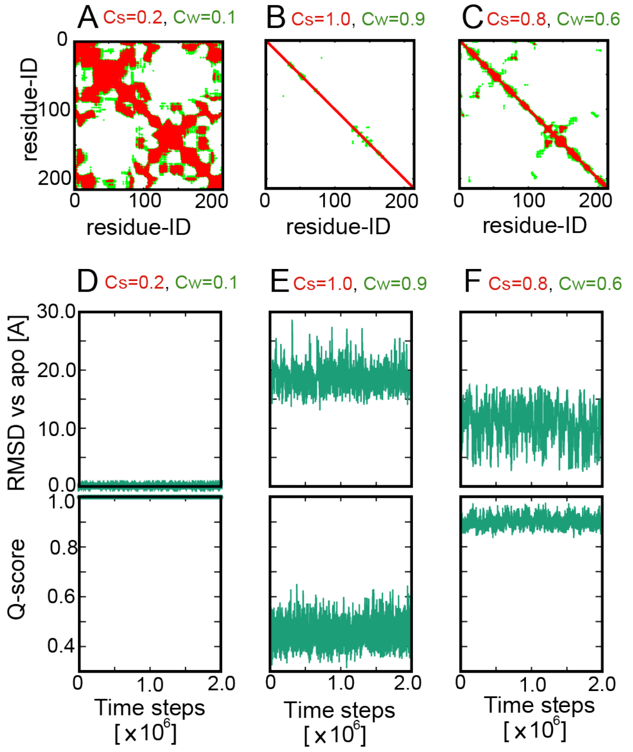

Figure S3. Color maps of the assigned interaction strengths K_{ij}: $(Ks, Kw) = (10, 1.0)$ of the adaptive CG-ENM based on DCCM for ADK with various (*Cs*, *Cw*) sets and the corresponding time

evolutions of RMSD from the apo-structure and Q-score based on apo-structure. (A, B, C) Color maps of the assigned interaction strengths in the residue pair plane with various (*Cs*, *Cw*) = (0.2, 0.1), (1.0, 0.9), (0.8, 0.6). The strong springs (*Ks* = 10) are assigned for residue pairs at red points, whereas the weak springs (*Kw* = 1) are assigned for residue pairs at green points. (D, E, F) Time evolution of RMSD from the apo-structure and the Q-score (apo) for corresponding parameters.

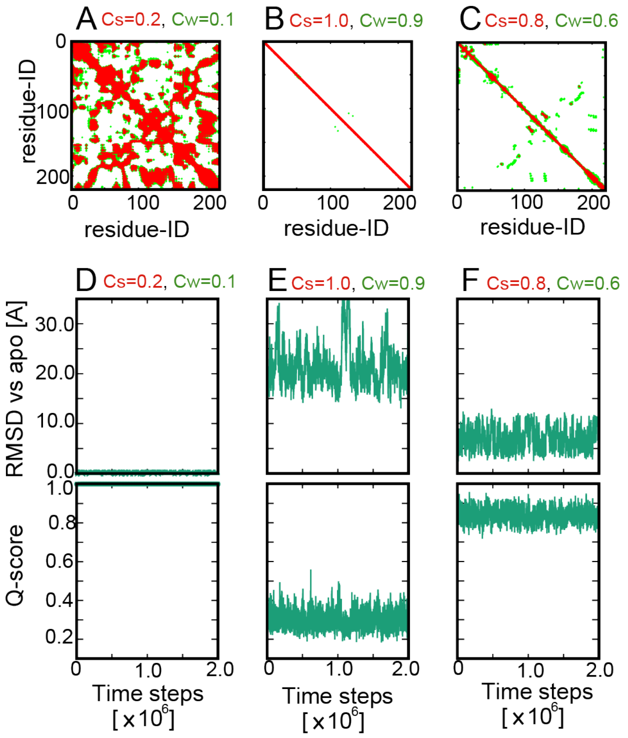

Figure S4. Color maps of the assigned interaction strengths K_{ij}: $(Ks, Kw) = (10, 8.0)$ of the adaptive CG-ENM based on DCCM for GBP with various (*Cs*, *Cw*) sets and the corresponding time evolutions of RMSD from the apo-structure and Q-score based on apo-structure. (A, B, C) Color maps of the assigned interaction strengths in the residue pair plane with various (*Cs*, *Cw*) = (0.2, 0.1), (1.0, 0.9), (0.8, 0.6). The strong springs (*Ks* = 10) are assigned for residue pairs at the red points, whereas the weak springs (*Kw* = 8) are assigned for residue pairs at the green points. (D, E, F) Time evolutions of RMSD from the apo-structure and the Q-score (apo) for corresponding parameters.

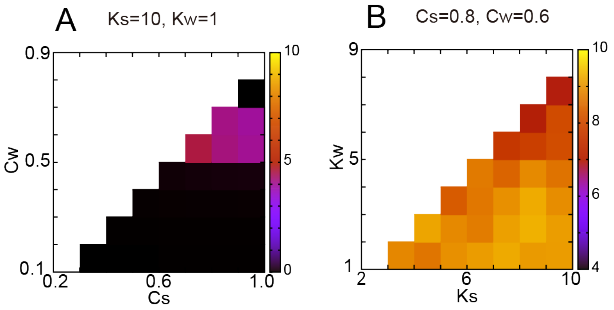

Figure S5. Contour plots of the score function: F_{BO} for ADK. (A) Contour plot of the score function on the (Cs, Cw) plane under the specific condition: (*Ks*, *Kw*) = (10, 1). (B) Contour plot of the score function on the (*Ks*, *Kw*) plane under the specific condition: (*Cs*, *Cw*) = (0.8, 0.6).

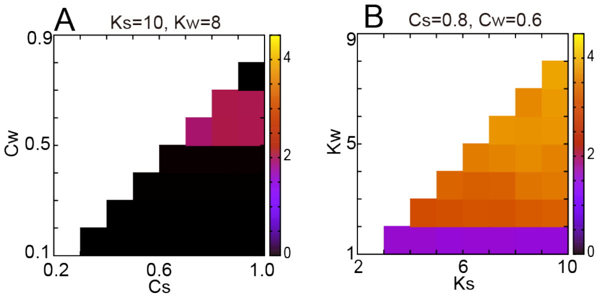

Figure S6. Contour plots of the score function: F_{BO} for GBP. (A) Contour plot of the score function on the (*Cs*, *Cw*) plane under the specific condition: (*Ks*, *Kw*) = (10, 8). (B) Contour plot of the score function on the (*Ks*, *Kw*) plane under the specific condition: (*Cs*, *Cw*) = (0.8, 0.6).

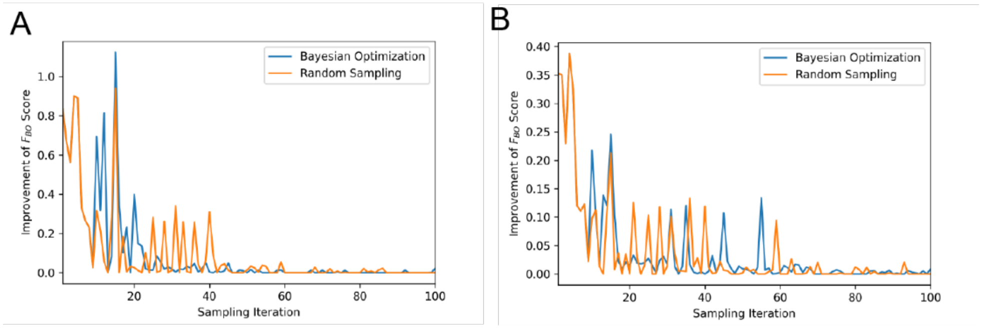

Figure S7. Sampling iteration number dependence of the averaged improvement score $\langle F_{BO}^{(i)} - F_{BO}^{(i-1)} \rangle$ over 30 trials in exploring suitable parameter sets by Bayesian optimization (BO) and random sampling (RS). (A) Result for ADK; (B) result for GBP. The blue and orange lines in each panel correspond to the average improvement score by BO and RS, respectively. The condition, such as the initially selected parameter set, is the same as for Figure 3.

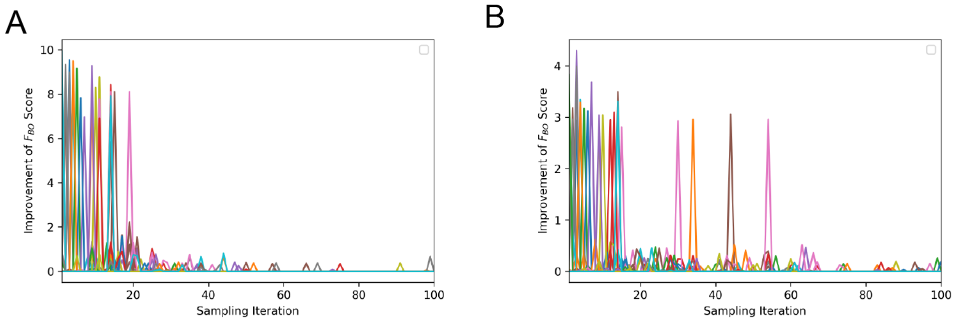

Figure S8. Sampling iteration number dependence of the raw improvement of score $F_{BO}^{(i)} - F_{BO}^{(i-1)}$ for every 30 trials in exploring suitable parameter sets by Bayesian optimization (BO). (A) Result for ADK; (B) result for GBP. In each panel, the different colored lines correspond to the result for different trials by BO. The condition, such as the initially selected parameter set, is the same as for Figure 3 and Figure S7.

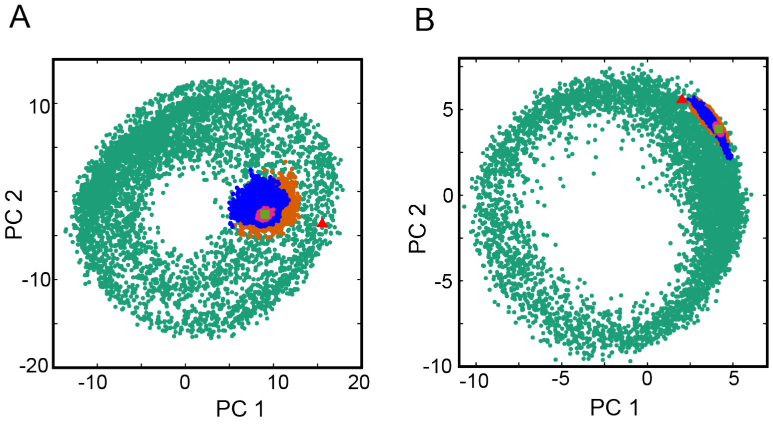

Figure S9. Comparison of the new adaptive-ENM with conventional ENM and AA-MD for ADK (A) and GBP (B). For adaptive CG-ENM, another suitable parameter set was utilized: (*Ks*, *Kw*, *Cs*, *Cw*) = (7.0, 5.0, 0.9, 0.6) and (8.0, 6.0, 0.7, 0.6) for ADK and GBP, respectively. Sampling points for adaptive-ENM, ENM(TREMD), and AA-MD (50 ns and 1 μs) are colored green, magenta, blue, and orange, respectively. Reference (apo) and target (holo) structures are colored light green (square) and red (triangle). (Sampling points are ploted in PCA plane, of which the PC1 and PC2 axes are defined by the ensemble via adaptive CG-MD.)

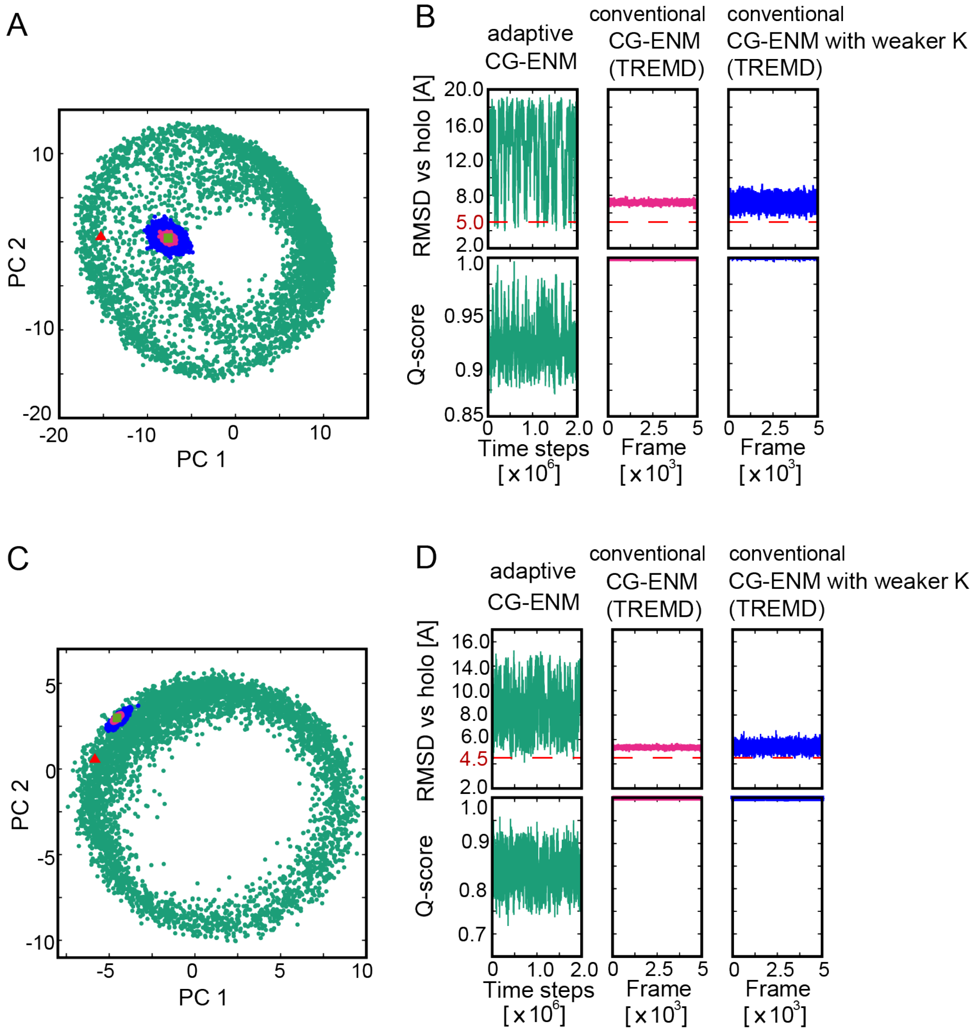

Figure S10. Sampling performance comparison between the new adaptive CG-ENM and conventional CG-ENM with default and weaker spring. (A and C) Comparison of structural ensembles sampled by adaptive CG-ENM and conventional CG-ENM with default spring and 10-fold weaker spring in the PCA plane, of which PC-12 axes are defined by the ensemble by adaptive CG-MD for ADK with (*Ks*, *Kw*, *Cs*, *Cw*) = (8.0, 7.0, 0.8, 0.6) and GBP with (*Ks*, *Kw*, *Cs*, *Cw*) = (10.0, 8.0, 0.8, 0.6), respectively. Sampling points for adaptive CG-ENM, conventional CG-ENM(TREMD) with default spring, and conventional CG-ENM(TREMD) with weaker spring are colored green, magenta, and blue, respectively. Reference (apo) and target (holo) structures are colored light green (square) and red (triangle). (B and D) Time evolution of RMSD vs target: holo (upper panels) and Q-score based on apo-structure (lower panels) for adaptive-ENM, conventional ENM with default spring, and with 10-fold weaker spring for ADK and GBP, respectively.

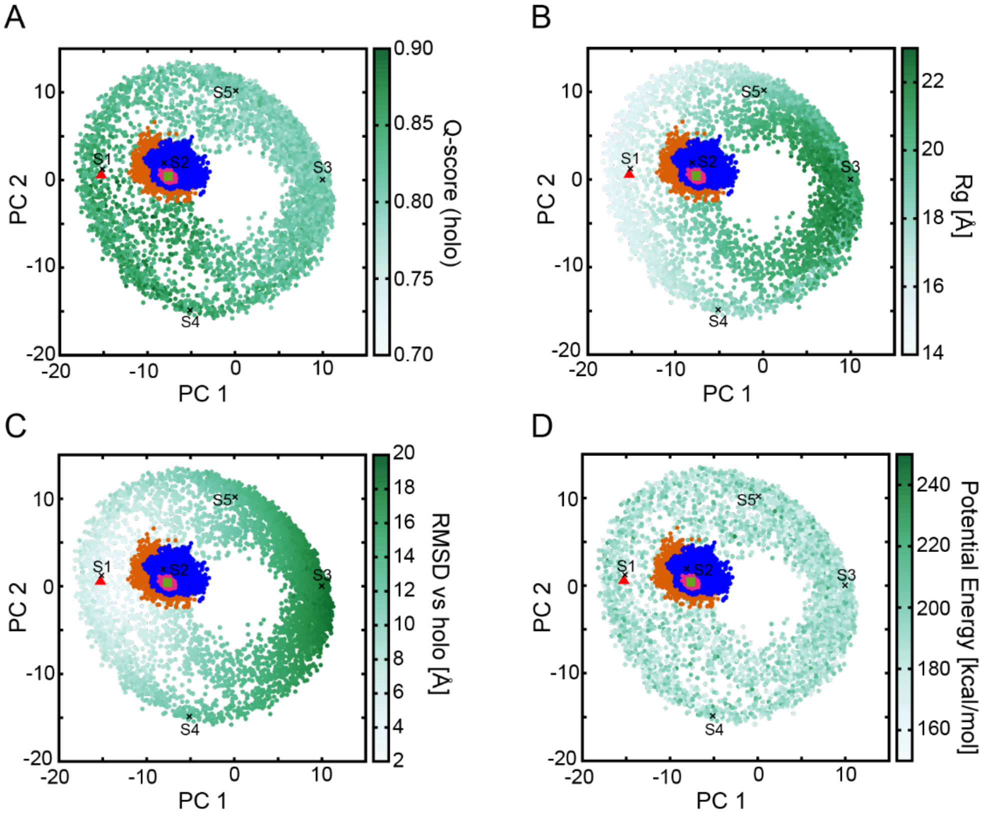

Figure S11. Comparison of structural ensembles sampled by adaptive-ENM, conventional ENM, and AA-MD (50 ns and 1 *μ*s) in the PCA plane, of which the PC-1 and PC-2 axes are defined by the ensemble via adaptive CG-MD for ADK. As same in Figure 5A, sampling points for adaptive-ENM with (*Ks*, *Kw*, *Cs*, *Cw*) = (8.0, 7.0, 0.8, 0.6), ENM(TREMD), and AA-MD (50 ns and 1 *μ*s) are colored green, magenta, blue, and orange, respectively. Green gradation for sampling points of adaptive-ENM depend on Q-score (holo) in panel (A), radius of gyration Rg in panel (B), RMSD vs holo in panel (C), and potential energy in panel (D), respectively.

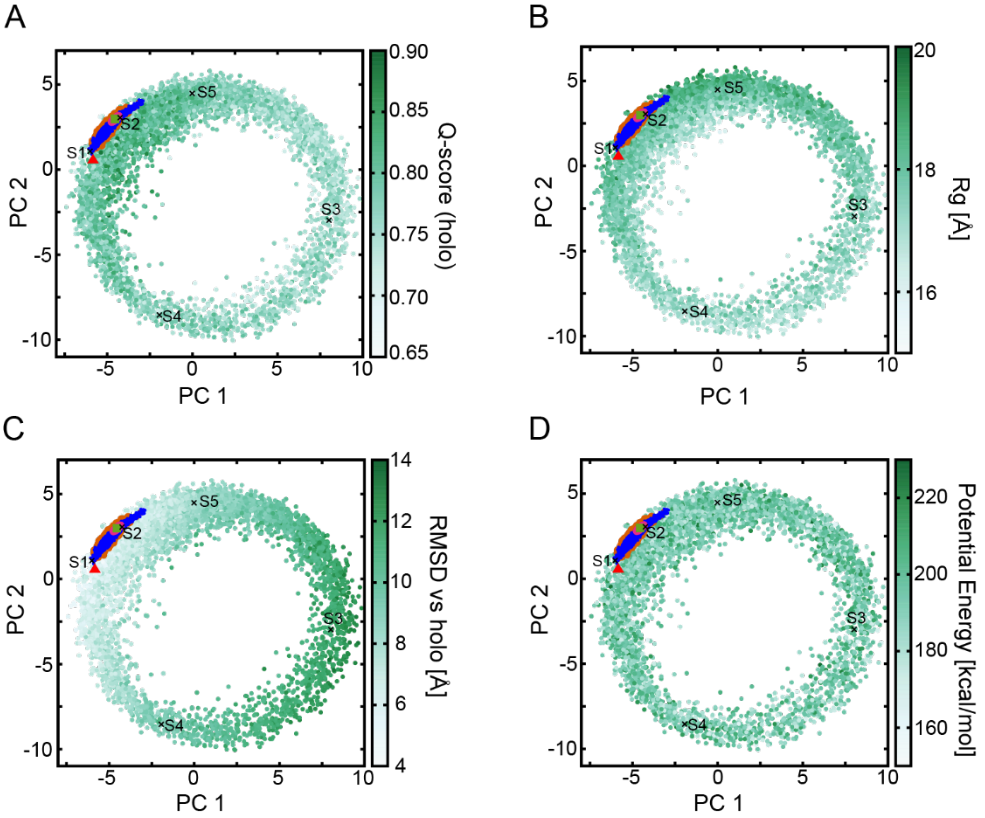

Figure S12. Comparison of structural ensembles sampled by adaptive-ENM, conventional ENM, and AA-MD (50 ns and 1 μs) in the PCA plane, of which the PC-1 and PC-2 axes are defined by the ensemble via adaptive CG-MD for GBP. As same in Figure 6A, sampling points for adaptive-ENM with (Ks, Kw, Cs, Cw) = (10.0, 8.0, 0.8, 0.6), ENM(TREMD), AA-MD (50 ns and 1μs) are colored green, magenta, blue, and orange, respectively. Green gradation for sampling points of adaptive-ENM depend on Q-score (holo) in panel (A), the radius of gyration Rg in panel (B), RMSD vs holo in panel (C), and Potential energy in Panel (D), respectively.

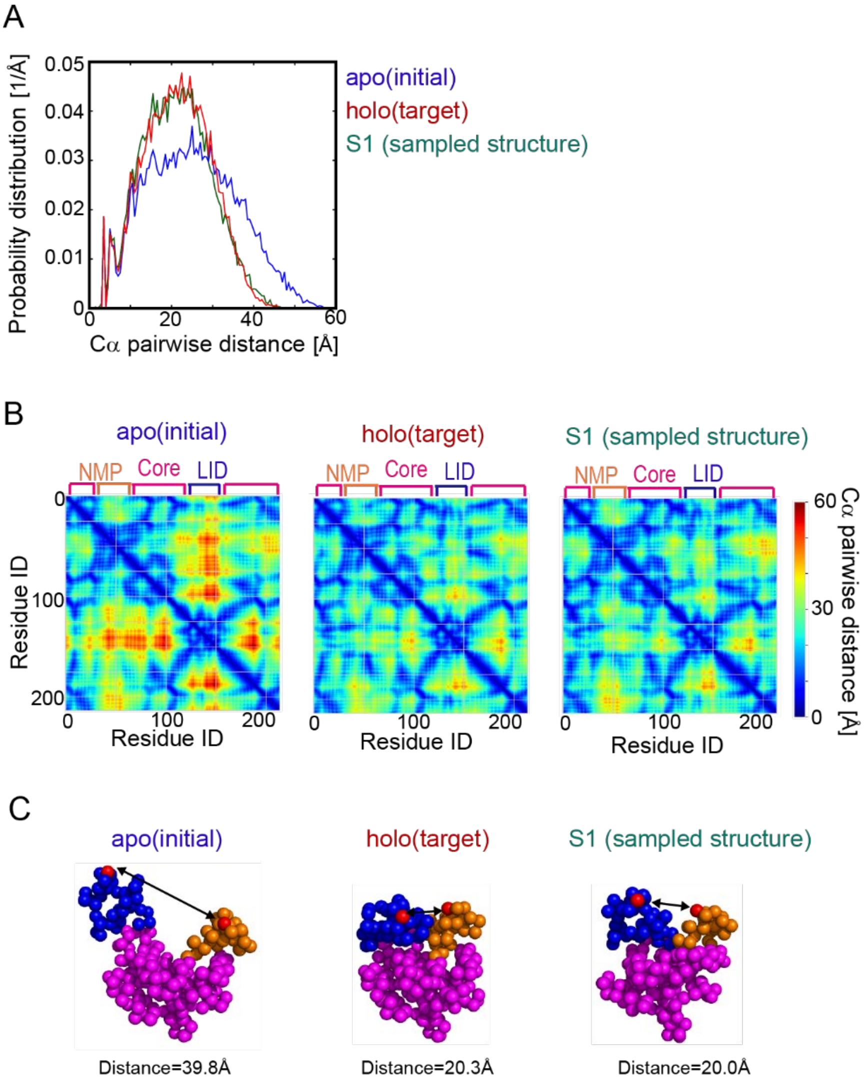

Figure S13. Detailed investigation of representative structure S1 sampled by adaptive CG-ENM and comparison with apo (initial) and holo (target) X-crystal structure for ADK. Model S1, of which RMSD to holo is significantly smaller among structural ensemble sampled by adaptive CG-ENM, is the same as the one shown in Figure 5. (A) Probability distributions of Cα pairwise distance for the whole of ADK in apo (initial), holo (target), and S1. (B) Cα distance matrix for all residue pairs. (C) Snapshot of apo, holo, and S1 structure. Black arrows indicate the Cα distance between representative residues resid:40 in NMP-domain and resid:149 in LID-domain.

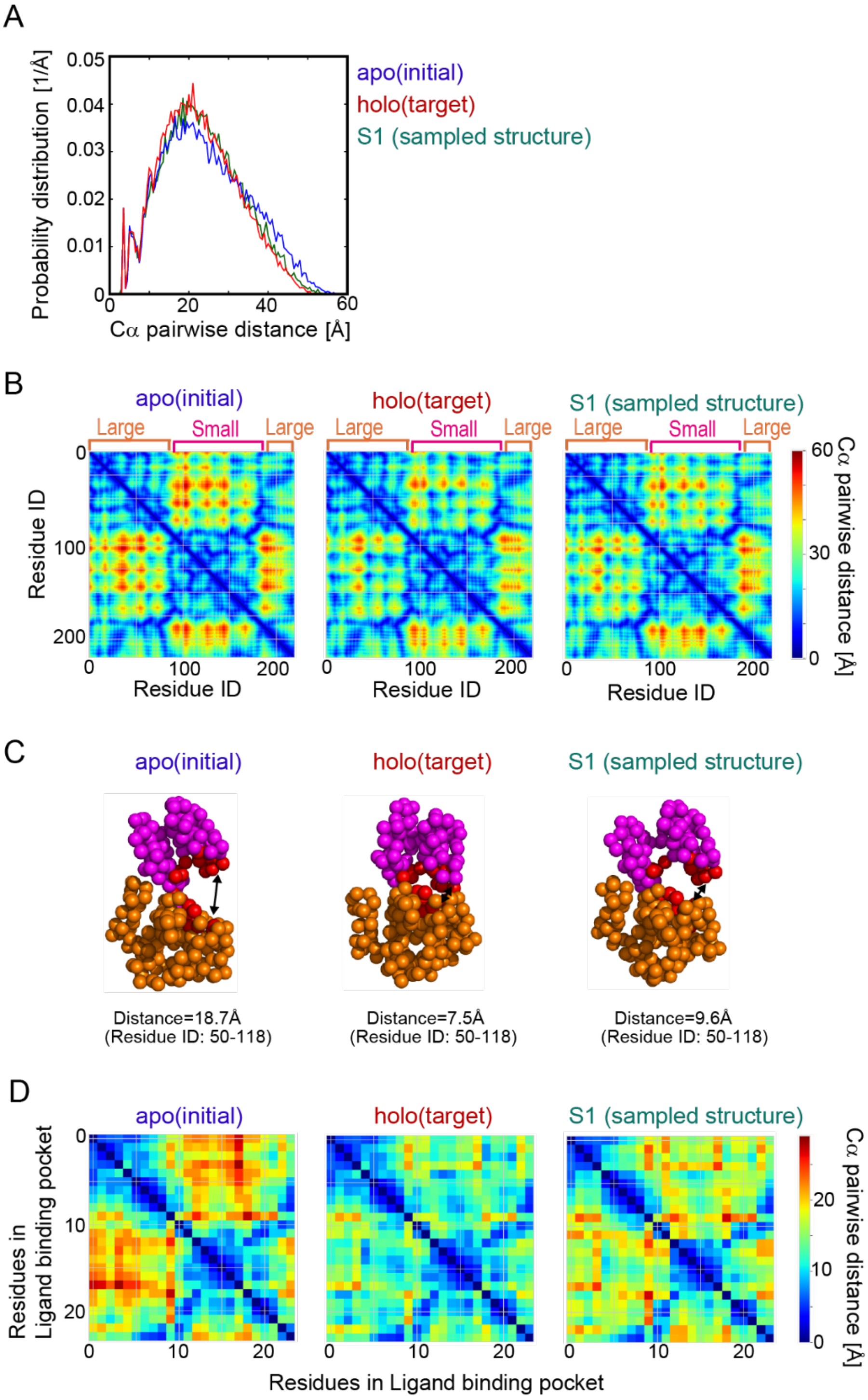

Figure S14. **Detailed investigation of representative structure S1 sampled by adaptive CG-ENM and comparison with apo(initial) and holo(target) X-crystal structure for GBP.** Model S1, of which RMSD to holo is significantly smaller among structural ensemble sampled by adaptive CG-ENM, is the same as the one shown in Figure 6. (A) Probability distributions of Cα pairwise distance for the whole of GBP in apo(initial), holo(target), and S1. (B) Cα distance matrix for all residue pairs. (C) Snapshot of apo, holo, and S1 structure. Black arrows indicate the Cα distance between two representative (ligand-binding) residues resid:50 in Large-domain and resid:118 in Small-domain. (D) A limited Cα distance matrix only for ligand-binding-pocket residue pairs. The residues of the Ligand-binding pocket are colored red in panel (C).

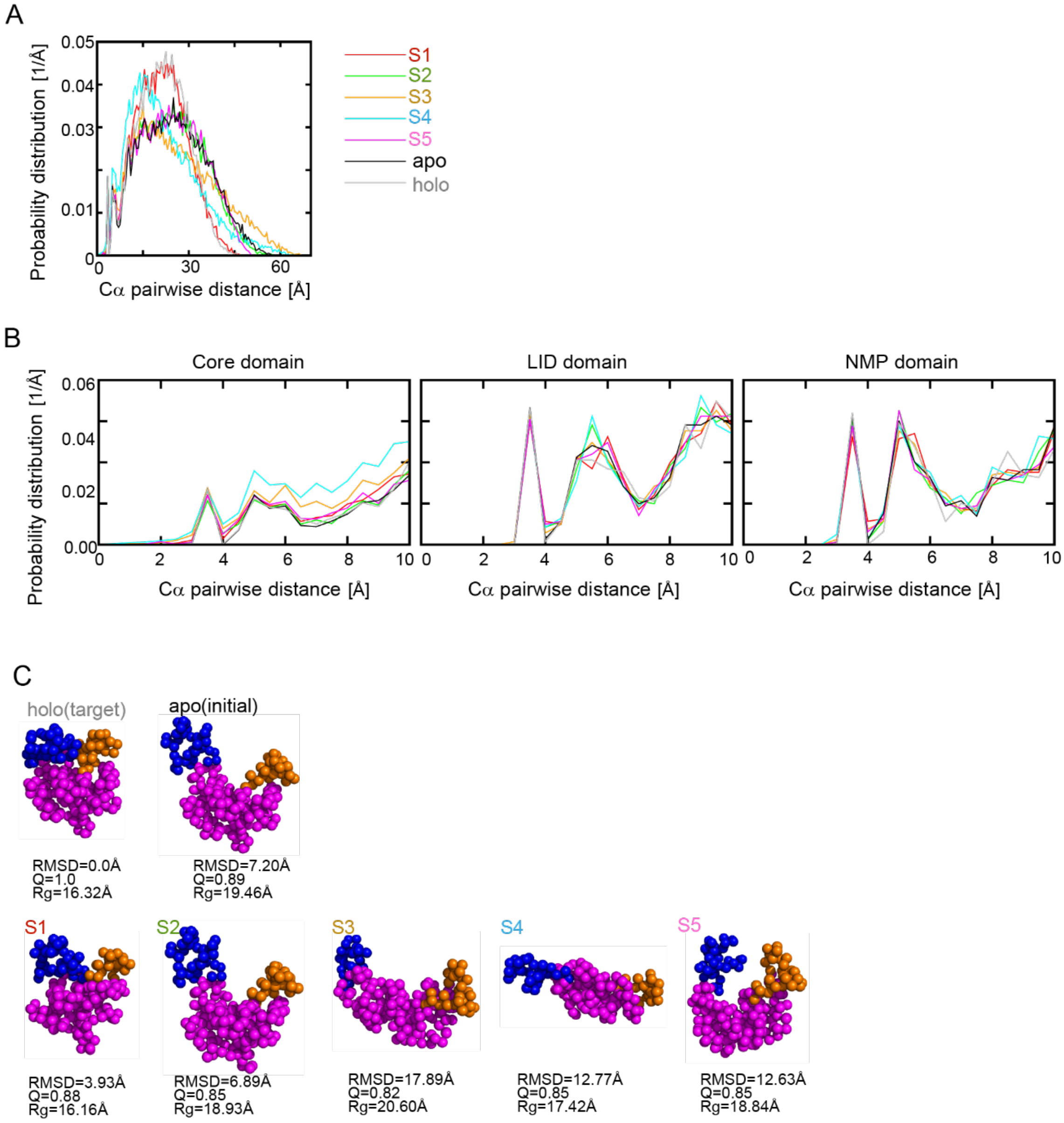

Figure S15. **Detailed representative structures S1–5 sampled by adaptive CG-ENM** for ADK. (A) Probability distributions of Cα pairwise distance for the whole of ADK in apo(initial), holo(target), and representative models S1-5. (B) Probability distributions of Cα pairwise distance for each intra-domain of ADK: core, LID, and NMP domain. (C) Snapshot of apo, holo, and S1-5 model. RMSD vs holo and Q-score based on holo and Rg values are added to each snapshot.

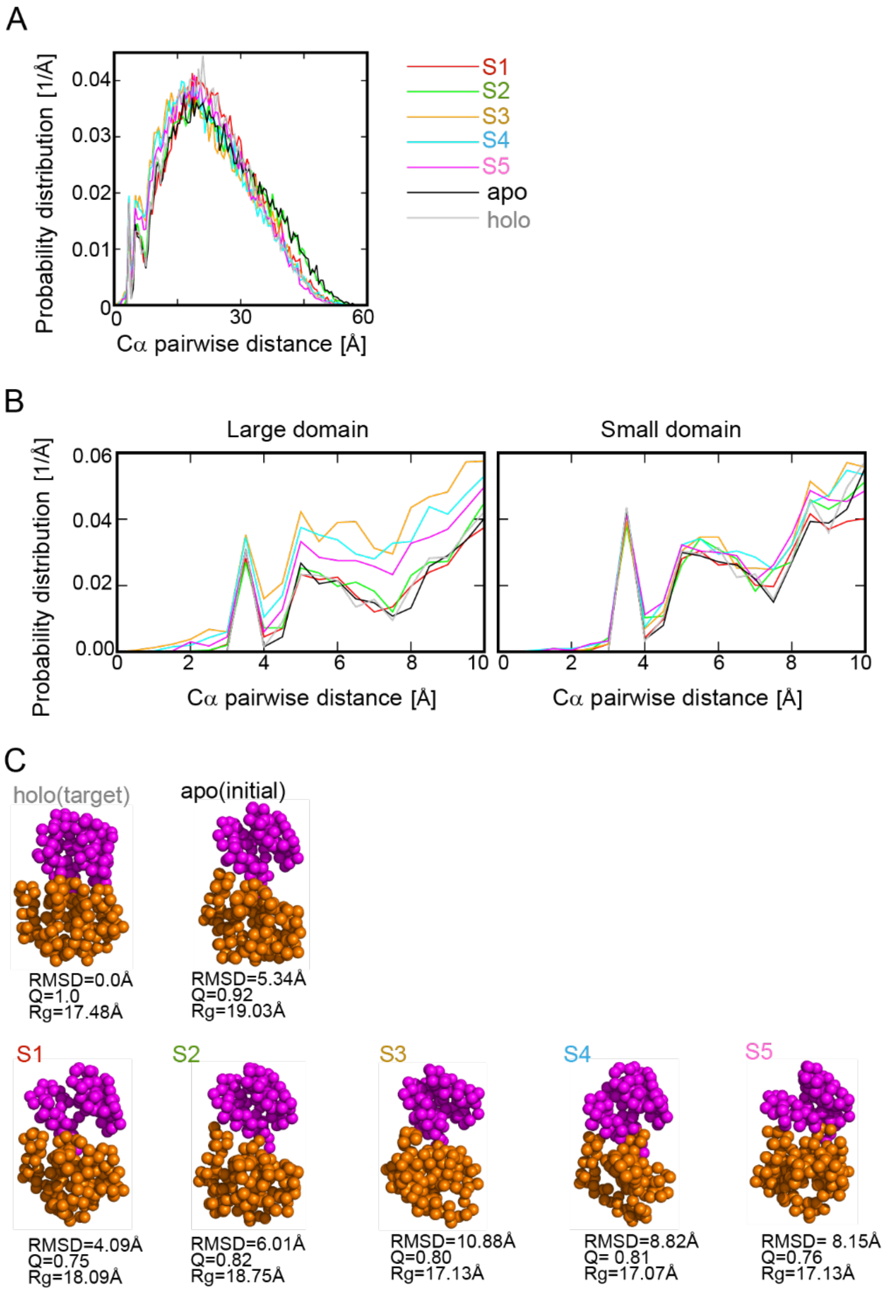

Figure S16. **Detailed representative structures S1–5 sampled by adaptive CG-ENM** for GBP. (A) Probability distributions of Cα pairwise distance for the whole of GBP in apo(initial), holo(target), and representative models S1-5. (B) Probability distributions of Cα pairwise distance for each intra-domain of GBP: Large and Small domain. (C) Snapshot of apo, holo, and S1-5 model. RMSD vs holo and Q-score based on holo and Rg values are added to each snapshot.

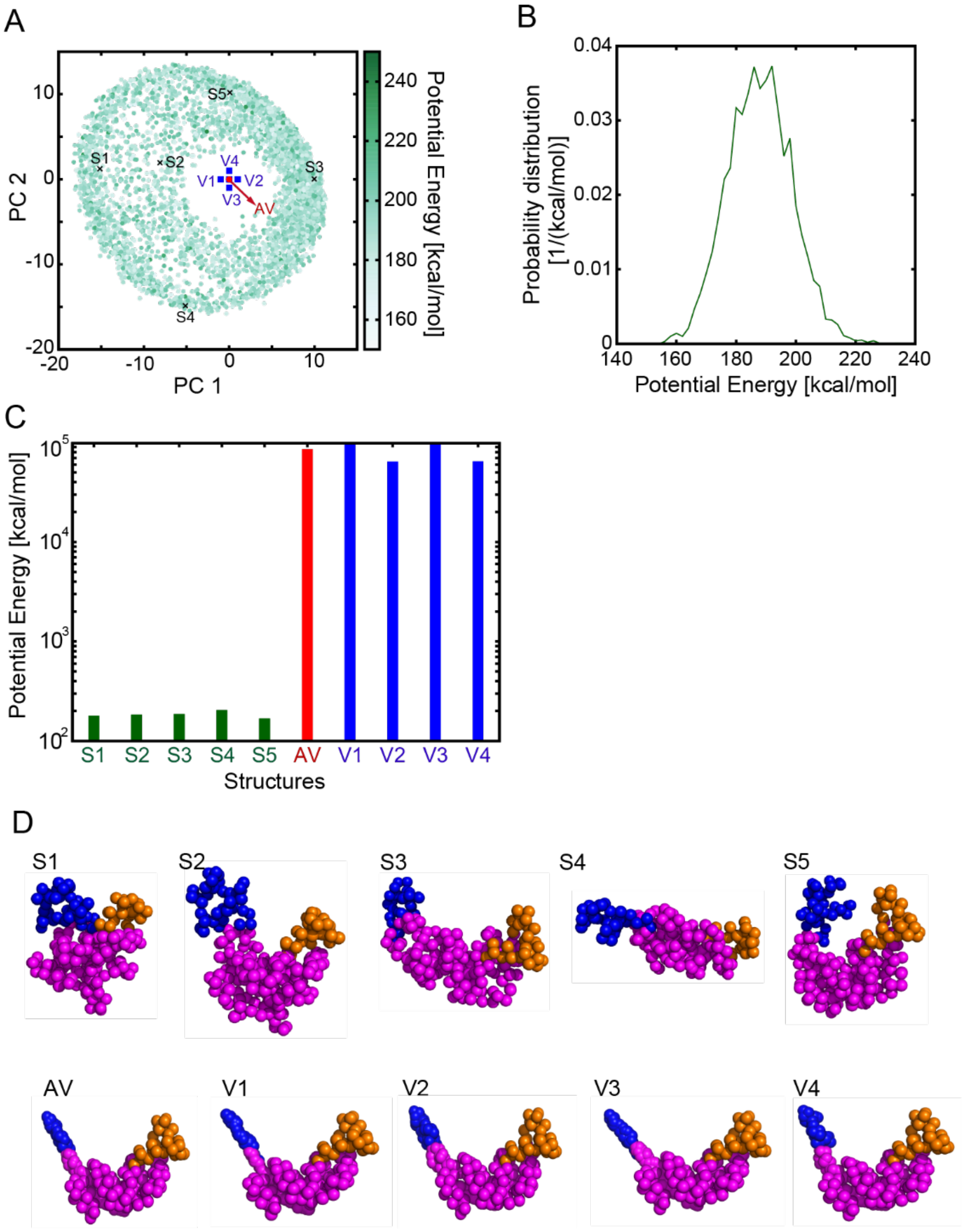

Figure S17. **Potential energy of ensemble structures sampled by adaptive CG-ENM and averaged structure "AV" and its neighborhood structures (V1, V2, V3, V4) around AV for ADK**. The structure ensemble is same as one shown in Figure 5. (A) The distributions of the potential energy of structures by adaptive CG-ENM in the PC1-2 plane. The point of the average structure AV and its neighborhood structures (V1, V2, V3, V4) are also plotted at (PC1, PC2) = {(0, 0), (-1, 0), (1, 0), (0, -1), (0, 1)} in the PC1-2 plane. (B) Probability distribution of potential energy for structural ensemble sampled by adaptive CG-ENM. (C) Comparison of potential energies for representative structures S1–5 sampled by adaptive CG-ENM with the ones for the modeled structures AV and V1–4. (D) Snapshot of representative sampled structures S1–5 and modeled structures AV and V1–4.

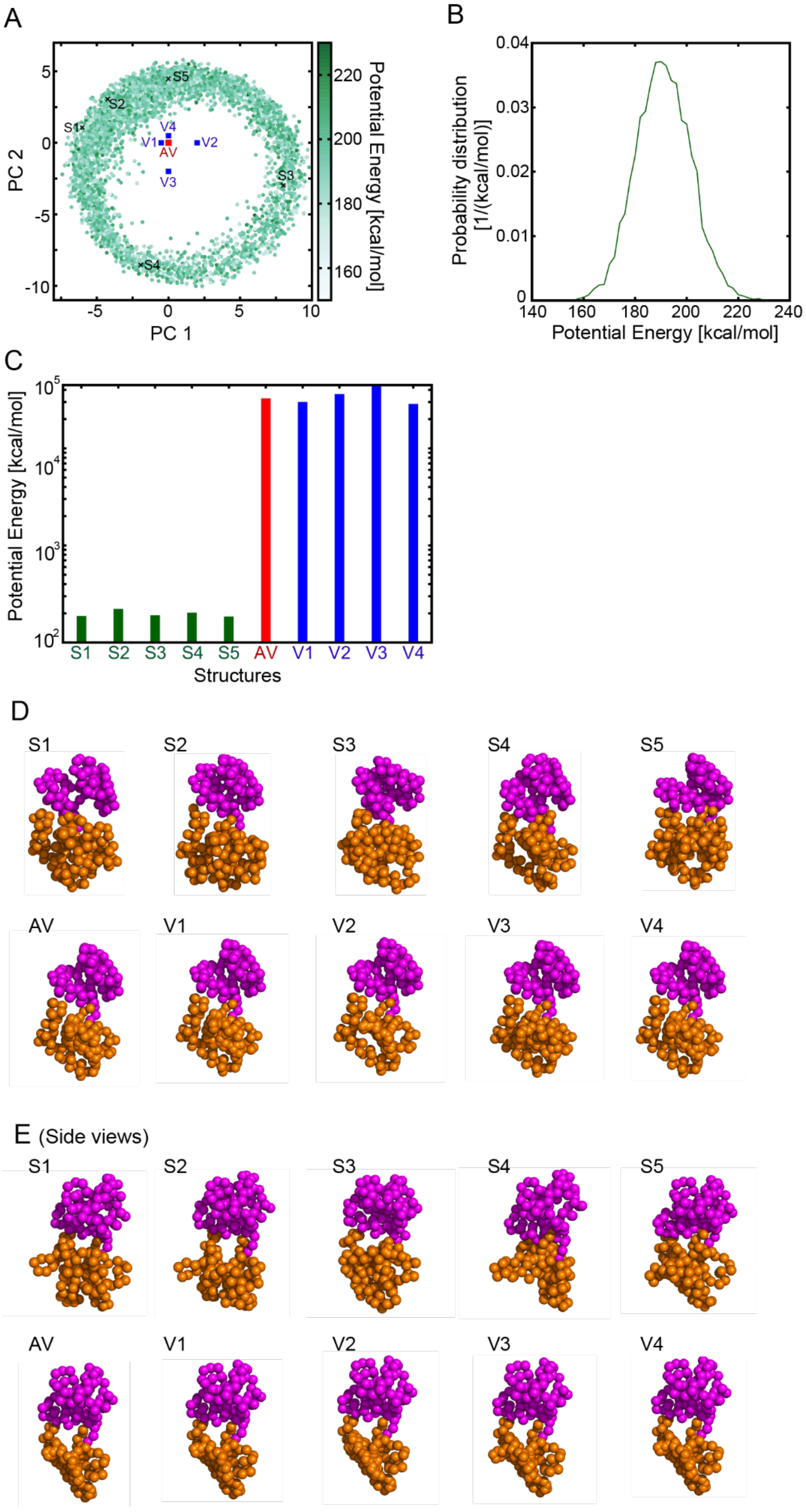
A
PC 2
PC 1
5
0
-5
-10
-5
0
5
10
S1
S2
S3
S4
S5
V1
V2
V3
V4
AV
Potential Energy [kcal/mol]
220
200
180
160
B
Probability distribution
[1/(kcal/mol)]
0.04
0.03
0.02
0.01
0
140
160
180
200
220
240
Potential Energy [kcal/mol]
C
Potential Energy [kcal/mol]
10^5
10^4
10^3
10^2
S1
S2
S3
S4
S5
AV
V1
V2
V3
V4
Structures
D
S1
S2
S3
S4
S5
AV
V1
V2
V3
V4
E (Side views)
S1
S2
S3
S4
S5
AV
V1
V2
V3
V4

Figure S18. **Potential energy of ensemble structures sampled by adaptive CG-ENM and averaged structure “AV” and its neighborhood structures (V1, V2, V3, V4) around AV for GBP.** The structure ensemble is the same as the one shown in Figure 6. (A) The distributions of the potential energy of structures by adaptive CG-ENM in the PC1-2 plane. The point of the average structure AV and its neighborhood structures (V1, V2, V3, V4) are also plotted at (PC1, PC2) = {(0, 0), (-0.5, 0), (2, 0), (0, -2), (0, 0.5)} in the PC1-2 plane. (B) Probability distribution of potential energy for structural ensemble sampled by adaptive CG-ENM. (C) Comparison of potential energies for representative structures S1–5 sampled by adaptive CG-ENM with the ones for modeled structures AV and V1–4. (D) Snapshot of representative sampled structures S1–5 and modeled structures AV and V1–4. (E) Sideview of corresponding snapshots shown in panel-D.

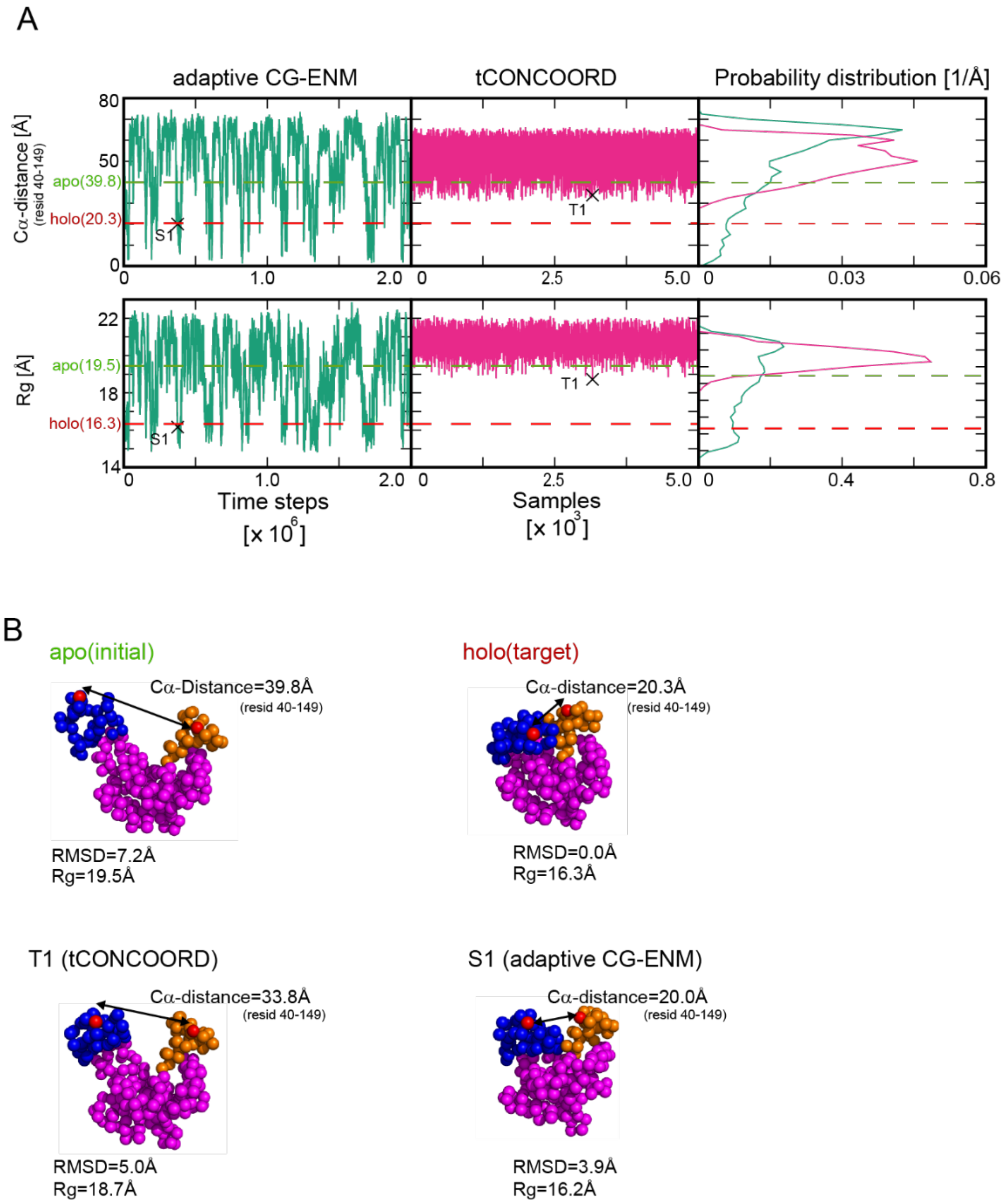

Figure S19. Sampling performance comparison for ADK between the new adaptive CG-ENM and tCONCOORD without using the Rg of the target (holo) as a constraint. (A) Time series for adaptive CG-ENM, sample-id dependence for tCONCOORD, and the corresponding probability distributions of Cα-distance between representative residues (residues 40–149) and radius of gyration Rg. The green and red dashed lines stand for the value of Cα-distance and Rg in the apo and holo state, respectively. Sampling with adaptive CG-ENM is conducted by (*Ks*, *Kw*, *Cs*, *Cw*) = (8.0, 7.0, 0.8, 0.6), and sampling with tCONCOORD is conducted by default set. (The structural ensemble for adaptive CG-ENM is the same as the one shown in Figure 5) (B) Snapshots of structure for apo, holo, and T1 with tCONCOORD, and S1 with adaptive

CG-ENM. RMSD of T1 to target (holo) is the smallest among the 5000 sampled structures by tCONCOORD, RMSD of S1 to target (holo) is significantly smaller among the 5000 sampled structures by adaptive CG-ENM. Sampling points for T1 and S1 are depicted in the time series of Cα-distance and Rg in panel A using a black “x”. RMSD vs holo, Cα-distance, and Rg values are added to each snapshot.

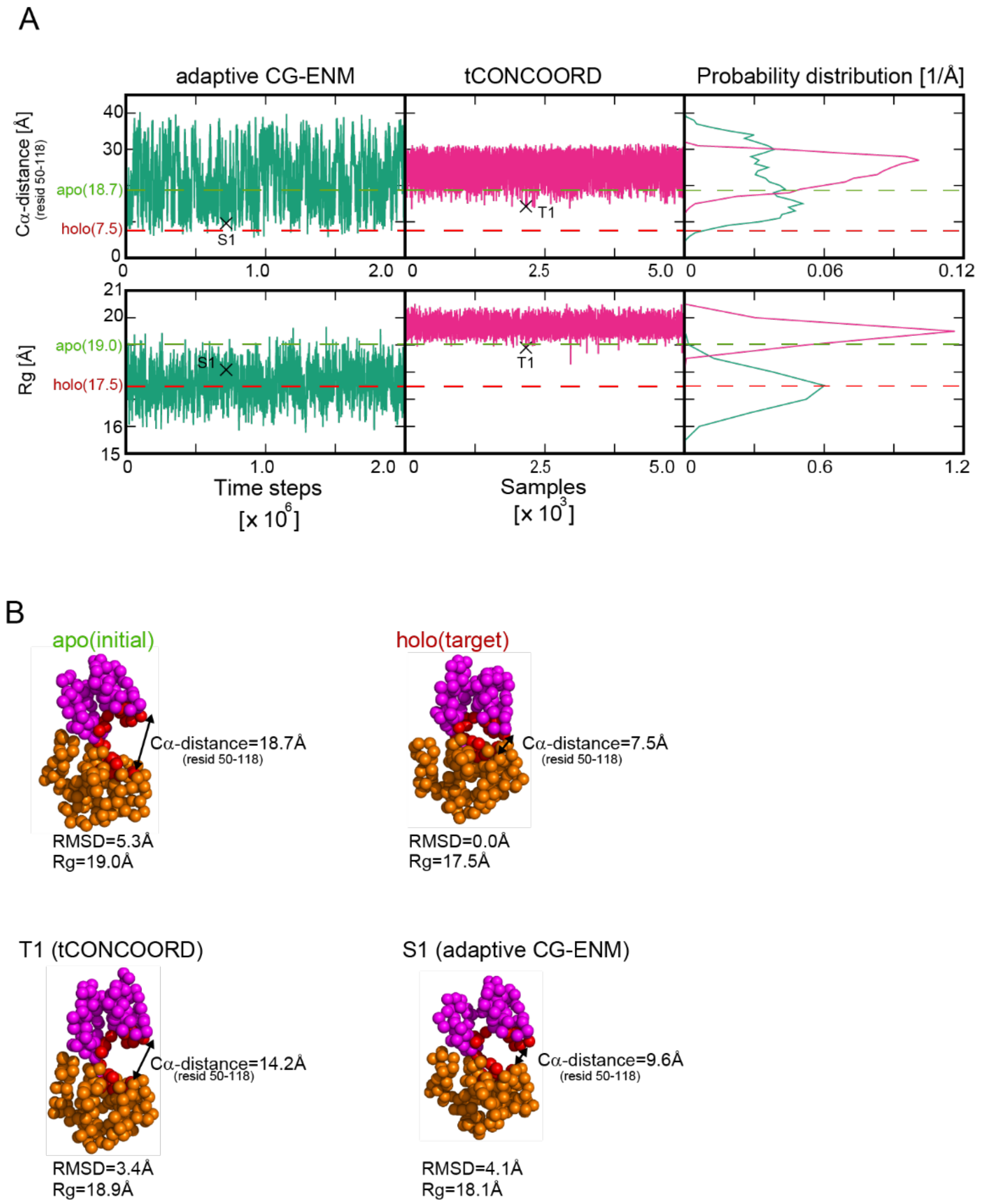

Figure S20. Sampling performance comparison for GBP between the new adaptive CG-ENM and tCONCOORD without using the Rg of the target (holo) as a constraint. (A) Time series for adaptive CG-ENM, sample-id dependence for tCONCOORD, and the corresponding probability distributions of Cα-distance between representative residues (residues 50–118) in binding-pocket and radius of

gyration Rg. The green and red dashed lines stand for the value of Cα-distance and Rg in apo ant holo state, respectively. Sampling with adaptive CG-ENM is conducted by (*Ks*, *Kw*, *Cs*, *Cw*) = (10.0, 8.0, 0.8, 0.6), and sampling with tCONCOORD is conducted by default set. (The structural ensemble for adaptive CG-ENM is the same as the one shown in Figure 6) (B) Snapshots of structure for apo, holo, and T1 with tCONCOORD, and S1 with adaptive CG-ENM. RMSD of T1 to target (holo) is the smallest among the 5000 sampled structures by tCONCOORD, RMSD of S1 to target (holo) is significantly smaller among the 5000 sampled structures by adaptive CG-ENM. Sampling points for T1 and S1 are depicted in time series of Cα-distance and Rg in panel A by using a black “x”. RMSD vs holo, Cα-distance, and Rg values are added to each snapshot.

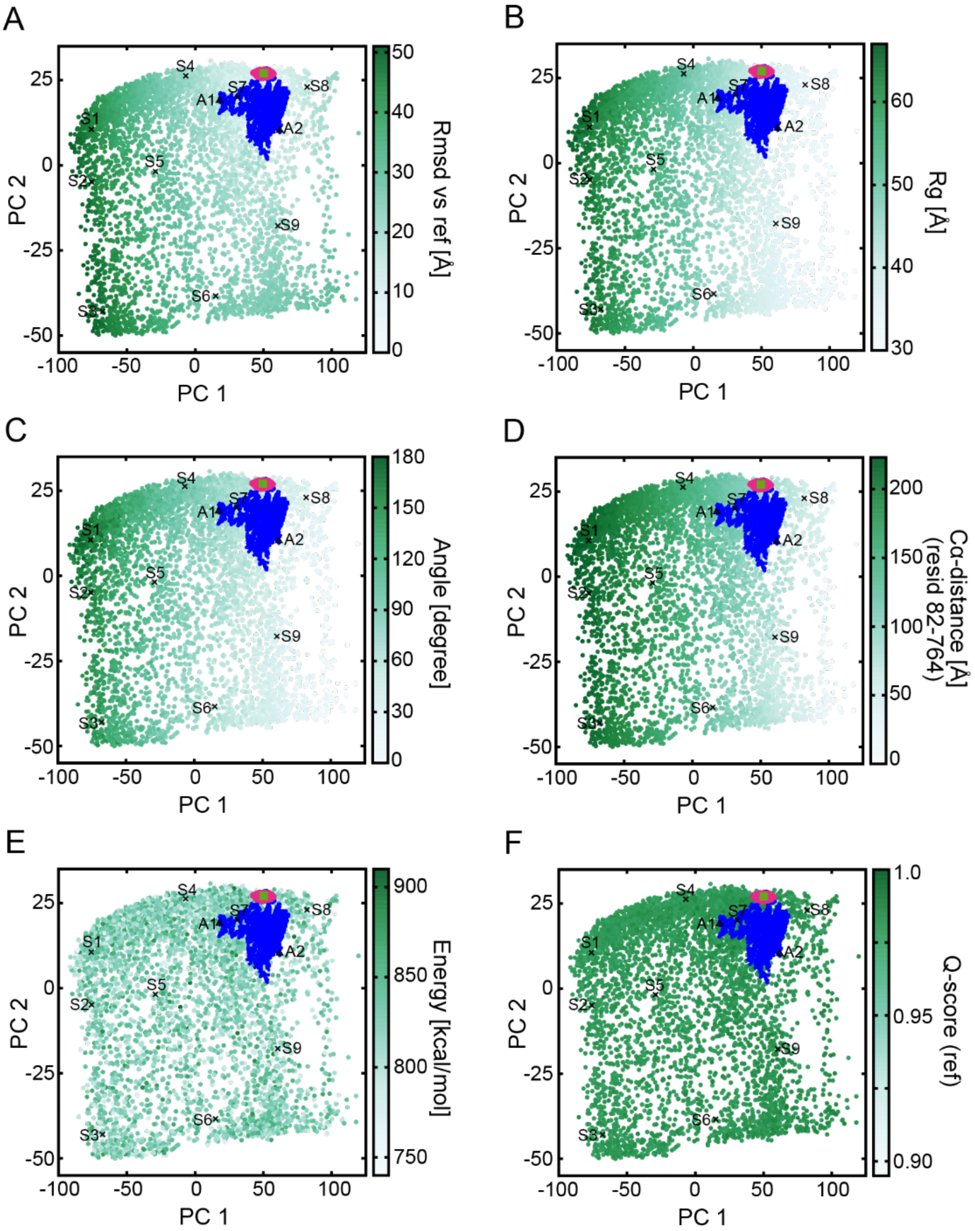

Figure S21. Comparison of structural ensembles sampled by adaptive CG-ENM, conventional CG-ENM, and AA-MD (50 ns) in the PCA plane, of which the PC-1 and PC-2 axes are defined by the ensemble via adaptive CG-MD for integrin. As same in Figure 9A, sampling points for adaptive CG-ENM with (*Ks*, *Kw*, *Cs*, *Cw*) = (7.0, 1.0, 1.0, 0.8), conventional CG-ENM(TREMD), and AA-MD (50 ns) are colored green, magenta, and blue, respectively. Green gradation for sampling points of adaptive-ENM depends on RMSD vs reference (inactive-state) in panel (A), the radius of gyration Rg in panel (B), Angle between two vectors that represent the direction of the long axis of Thigh-domain by $\boldsymbol{\gamma}_{571} - \boldsymbol{\gamma}_{470}$ and Calf-1 domain by $\boldsymbol{\gamma}_{619} - \boldsymbol{\gamma}_{602}$ in panel (C), Cα distance between representative residues (residues 82–764) in β-propeller and Calf-2 domain in panel (D), potential energy in panel (E), and Q-score (ref) based on inactive structure in panel (F).

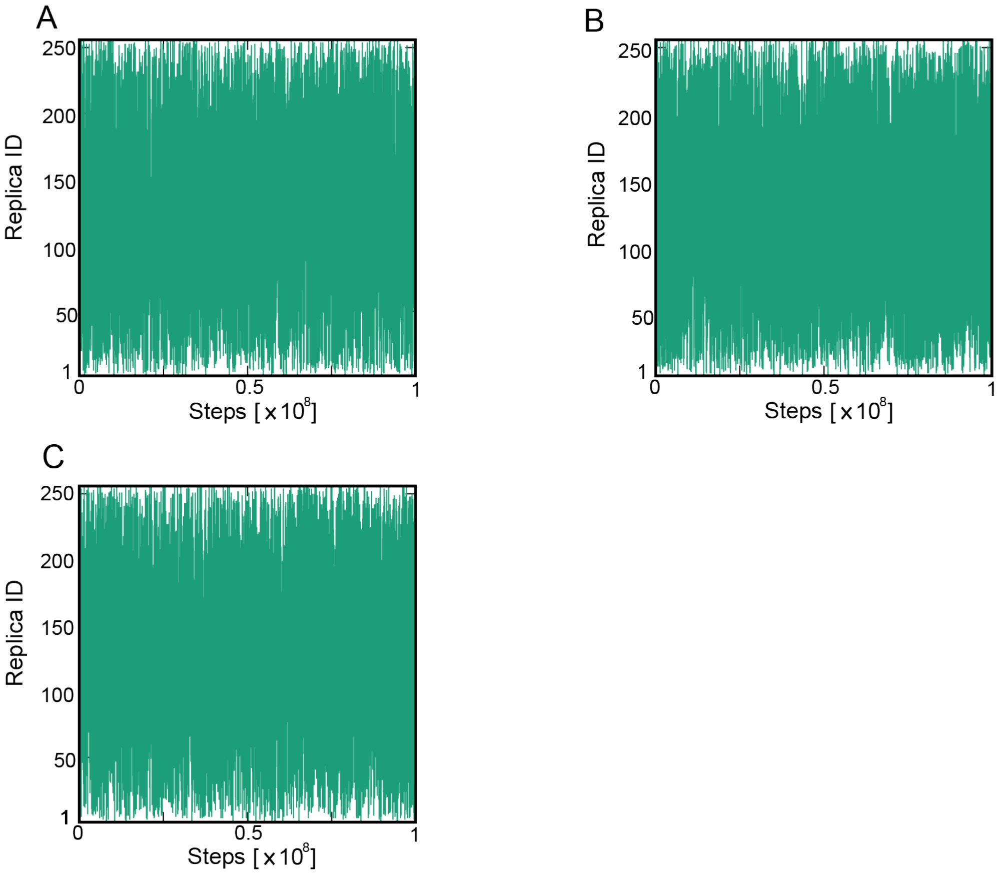

Figure S22. Time step dependence of replica exchange for ADK (A), GBP (B), and integrin (C). Replica ID at 300.0 K in conventional CG-ENM using TREMD simulation are plotted as the function of steps.

Supporting Information Table

Table S1. Surface areas of the bounding boxes for structural ensemble points explored by each model (adaptive CG-ENM, conventional CG-ENM by TREMD, and AA-MD of 1 μs) in the PC12-plane.

	Adaptive CG-ENM	Conventional CG-ENM by TREMD	AA-MD (1 μs)
ADK	1.00	2.27×10^{-3}	5.67×10^{-2}
GBP	1.00	3.58×10^{-4}	1.04×10^{-2}

The PC-12 axis is defined by the eigenvector of PCA based on trajectory that is generated by concatenating three trajectories with adaptive CG-ENM, conventional CG-ENM by TREMD, and

AA-MD of 1 μs. The surface areas of the bounding boxes for AA-MD and conventional CG-ENM by TREMD are normalized by that of adaptive CG-ENM. The trajectories of adaptive CG-ENM for ADK and GBP are reproduced via under-damped Langevin dynamics simulation with the respective suitable parameter (*Ks*, *Kw*, *Cs*, *Cw*) = (8.0, 7.0, 0.8, 0.6) and (10.0, 8.0, 0.8, 0.6) searched by BO. These parameter conditions are the same as used for **Table 1.**

Table S2. Correlation coefficient of the probability distribution of Cα pairwise distance and Cα distance matrix between sampled-model: S1 and (apo and holo) structures.

	Probability distribution of Cα pairwise-distance		Cα-distance matrix between all residue pairs		Limited Cα-distance matrix between ligand binding residue pairs	
	S1 vs apo	S1 vs holo	S1 vs apo	S1 vs holo	S1 vs apo	S1 vs holo
ADK	0.93	0.99	0.86	0.91	-	-
GBP	0.99	0.99	0.96	0.95	0.85	0.90

The correlation coefficient between S1 vs (apo, holo) structures are evaluated based on the probability distribution of Cα pairwise distance and Cα distance matrix shown in panels A and B of Figure S13 and S14 for ADK and GBP, respectively. In contrast, the correlation coefficient of the limited Cα-distance matrix between ligand binding residue pairs is evaluated based on the limited matrix shown in panel D of Figure S14.

REFERENCES in Supporting Information

(1) McCammon, J. A. Protein Dynamics. *Reports Prog. Phys.* **1984**, *47* (1), 1–46. https://doi.org/10.1088/0034-4885/47/1/001.

(2) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4* (3). https://doi.org/10.1021/ct700301q.

(3) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber Ff99SB Protein Force Field. *Proteins Struct. Funct. Bioinforma.* **2010**, *78* (8). https://doi.org/10.1002/prot.22711.

(4) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2).

https://doi.org/10.1063/1.445869.

(5) Nosé, S. A Molecular Dynamics Method for Simulations in the Canonical Ensemble. *Mol. Phys.* **1984**, *52* (2). https://doi.org/10.1080/00268978400101201.

(6) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81* (8). https://doi.org/10.1063/1.448118.

(7) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An $N \cdot \log(N)$ Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12). https://doi.org/10.1063/1.464397.

(8) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18* (12). https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.

(9) Seeliger, D.; de Groot, B. L. Conformational Transitions upon Ligand Binding: Holo-Structure Prediction from Apo Conformations. *PLoS Comput. Biol.* **2010**, *6* (1). https://doi.org/10.1371/journal.pcbi.1000634.

(10) Dokainish, H. M.; Sugita, Y. Exploring Large Domain Motions in Proteins Using Atomistic Molecular Dynamics with Enhanced Conformational Sampling. *Int. J. Mol. Sci.* **2020**, *22* (1). https://doi.org/10.3390/ijms22010270.

(11) Seeliger, D.; Haas, J.; de Groot, B. L. Geometry-Based Sampling of Conformational Transitions in Proteins. *Structure* **2007**, *15* (11). https://doi.org/10.1016/j.str.2007.09.017.

(12) Seeliger, D.; De Groot, B. L. TCONCOORD-GUI: Visually Supported Conformational Sampling of Bioactive Molecules. *J. Comput. Chem.* **2009**, *30* (7). https://doi.org/10.1002/jcc.21127.

(13) Björkman, A. J.; Mowbray, S. L. Multiple Open Forms of Ribose-Binding Protein Trace the Path of Its Conformational Change. *J. Mol. Biol.* **1998**, *279* (3), 651–664. https://doi.org/10.1006/jmbi.1998.1785.