Supplementary Information

Assessment of the utility of contact-based restraints in accelerating the prediction of protein structure using molecular dynamics simulations

Structure pre-processing for extended-state simulations

Structures were initially prepared in extended conformation using Maestro.[1] Prior to carrying out replica exchange molecular dynamics (REMD) simulations in vacuo, the force constant of the omega torsional potential for the protein backbone was increased by 10.0 kcal mol^{-1} in order to prevent cis-trans isomerization of peptide groups at high temperature. Distance restraints were implemented in Desmond[2] as flat-bottomed harmonic potentials with a spring constant of 1.0 kcal $(mol\ Å^2)^{-1}$ using the enhanced-sampling plugin. REMD simulations on Desmond were carried out at 16 exponentially separated temperatures ranging from 300 K to 700 K. Each of these simulations was run for 50 ns in the Nosé-Hoover NVT ensemble.[3–5] The most compact structure among the structures satisfying the maximal number of contacts was extracted from the 300 K trajectory of the REMD simulations and used as a starting point for simulated annealing simulation in the presence of solvent. The REMD phase resulted in compact conformations that satisfied the majority of restraints (Figure S1).

Proteins were solvated in water boxes large enough to allow for an initial distance of 24 Å between protein images. Sodium and chloride ions were added to the system to neutralize the protein charge and to set the salt concentration to 0.1 mol L^{-1}. Histidine residues were simulated in the N^{ε} protonated neutral state, while aspartate, glutamate, arginine, and lysine residues were simulated in their charged state. The CHARMM22* force field[6] with a TIP3P water model[7] was used to describe the systems. A cutoff radius of 10.8 Å was used to separate near and distant

electrostatic interactions, the latter being accounted for with the *k*-space Gaussian split Ewald method.[8] Tail corrections for Lennard-Jones interactions[9] were included in the virial but not in the energy calculations. The force constant of the omega torsion potential was reset to its original value. Simulated annealing simulations were carried out on Desmond for 40 ns with a temperature ramp-up from 300 K to 350 K and back to 300 K, as described in the main text (Methods). The final snapshot from the simulated annealing simulation was used as a starting point for all-atom production simulation on Anton.[10] The simulated annealing phase resulted in physically realistic structures, but with no significant improvement in the number of satisfied restraints or in the RMSD from the native conformation (Figure S2). In each case, the last snapshot from the simulated annealing simulation was taken as the initial system for a long simulated tempering simulation on Anton.

Secondary structure formation during pre-processing

In the case in which we restrained the ubiquitin structure with all 205 distance restraints (based on the full set of non-redundant contacts), we found significant secondary structure formation during the replica exchange and annealing phases (Figure S3). Specifically, the helix and the second hairpin were completely formed during these phases, while the first hairpin was partially formed. No secondary structure element was formed during the pre-processing stages in any of our other ubiquitin simulations.

Unrestrained simulated tempering simulations of ubiquitin from an extended conformation

To provide a baseline for assessing the extent of speedup of the folding rate that can be achieved by using distance restraint information, we performed four unrestrained simulated tempering simulations of ubiquitin starting from different conformations corresponding to the four most

compact conformations found during the REMD phase described above. Each unrestrained starting conformation was initially relaxed using the annealing protocol described above and then run for approximately 40–60 μs, for a total simulation time of about 200 μs. No transition to a native-like state was observed over this length of time (Figure S4), suggesting that the folding time of ubiquitin when using simulated tempering is a few hundred microseconds or longer. Indeed, we found that the only secondary structure element to be consistently formed within a few tens of microseconds of unrestrained simulation is hairpin 1, consistent with observations from simulations performed close to the melting temperature.[11]

Implementation of dihedral restraints based on secondary structural information

As described in the main text (Methods), Stride[12] was used to calculate the secondary structure from the experimentally determined three-dimensional structure, and Concord[13] was used to predict the secondary structure from sequence. A three-letter secondary structure alphabet was employed, consisting of helix (H), β strand (E) and random coil (C). In simulations performed with restraints based on the experimental secondary structure, backbone dihedral angles in helices and β strands were restrained with a torsional spring constant of 1.0 kcal mol^{-1}. Backbone dihedral angles in coil residues were not restrained. In simulations performed with restraints based on predicted secondary structure, the restraint spring constant k in helices and β strands was proportional to the confidence level of the prediction, c, which ranges between 0 and 9 for Concord predictions, as $k = 1.0\ (c / 9)$ kcal mol^{-1}. As in the exact secondary structural case, coil residues were not subjected to any dihedral angle restraints, regardless of their Concord confidence level. In all cases in which dihedral angle restraints were applied, these restraints were eventually removed after secondary structure formation.

Kinetic traps found in restrained simulations

The majority of our distance-restrained ubiquitin simulations culminated in kinetically trapped conformations that were close to the native conformation in terms of RMSD. We characterized the heterogeneity of these kinetically trapped conformations by carrying out all-against-all RMSD calculations between distinct simulations. Figure S6 shows histograms of all-against-all RMSD values between trajectory segments taken from two different simulations. The relevant trajectory segment for each simulation was selected as the portion of the trajectory that corresponded to the kinetic trap. These histograms, coupled with RMSD values from the native state (Figures 2 and 3 of the main text) suggest that the four 15-restraint simulations converge to the same kinetic trap, while '31_2' and '62_3' clearly correspond to different kinetic traps. Further analysis of hydrogen bonds (Figure S7) shows that the kinetic trap of the 15-restraint simulations corresponds to a register shift in hairpin 2, whereas the hydrogen bonding network of hairpin 2 is completely disrupted in the '62_2' simulation. The '62_3' simulation, on the other hand, converges to a state in which hairpin 1 suffers a register shift. The native hydrogen bonding patterns in the two β hairpins are not disrupted in the '31_2' simulation. Instead, the simulation is locked in a state in which loop 2 adopts a non-native conformation (Figure S7G).

Identification of native-like states from the Tc684 restrained simulation trajectory

In the restrained simulation trajectory of Tc684 starting from an extended conformation, we identified snapshots that satisfied all of the eight distance restraints supplied by CASP, and clustered these snapshots using *K*-means clustering for different values of *K*, as described (Methods). We identified the largest cluster for each value of *K* and selected the snapshot in the largest cluster that had the smallest average RMSD from other snapshots in the cluster. The RMSD of this "centroid" frame from the native conformation of ubiquitin is reported in Table S1

for different values of *K*. As is clear from the table, this simple scoring method robustly identifies frames in the native ensemble.

Other protocols for implementation of distance restraints

The temperature-based simulated tempering protocol employed in this work does not prevent protein structures from being stuck in kinetic traps. We thus ran some of the ubiquitin simulations with two other protocols: (a) one in which the distance restraints are weakened at high temperature, and (b) another in which the restraint force is constant at large distances. In protocol (a), the restraint potential is simultaneously tempered with the temperature, so that the potential is a linear function of temperature (note that the temperature rungs themselves are exponentially separated), attaining its maximum value of 0.05 kcal (mol Å^2)$^{-1}$ at the lowest temperature of 300 K and dropping to 0 at the highest temperature of 420 K. In protocol (b), the restraint potential function is chosen as

$$V(r) = \begin{cases} 0, & 0 \le r \le d\,, \\ \frac{1}{2}k(r-d)^2, & d < r \le d+2, \\ 2k(r-d-1), & d+2 < r, \end{cases}$$

where *d* is the restraint distance (10 Å for ubiquitin, 8 Å for the CASP targets), and *k* is the spring constant for the quadratic part of the potential, which is confined to a region within 2 Å of the restraint distance. The potential beyond this region is linear (constant force), and both the potential and the force are continuous everywhere.

We found protocol (a) to be difficult to implement in practice because of stability issues: Simulated tempering weights computed using averaged energies from short MD simulations consistently led to highly skewed occupancies of the temperature/potential ladder over short timescales (typically <1 μs), leading to unreasonably long round-trip times for transitions in temperature/potential space. Protocol (b), on the other hand, appears more promising. In ubiquitin simulations with 15 distance restraints, the use of the linear potential led to adoption of

the native conformation in two out of three cases (Figure S8). Due to the limited statistics, however, it is unclear if there is a significant difference in folding timescales relative to the harmonic restraint potential.

References

1. Maestro, version 9.8 (2014) Schrödinger, LLC, New York.

2. Bowers KJ, Chow E, Xu H, Dror RO, Eastwood MP, Gregersen BA, Klepeis JL, Kolossváry I, Moraes MA, Sacerdoti FD, Salmon JK, Shan Y, Shaw DE (2006) Scalable algorithms for molecular dynamics simulations on commodity clusters. Proceedings of the ACM/IEEE Conference on Supercomputing (SC06). New York: ACM.

3. Nosé S (1984) A unified formulation of the constant temperature molecular dynamics methods. J Chem Phys 81(1):511–519.

4. Hoover WG (1985) Canonical dynamics: Equilibrium phase-space distributions. Phys Rev A 31(3):1695–1697.

5. Martyna GJ, Klein ML, Tuckerman M (1992) Nosé-Hoover chains: The canonical ensemble via continuous dynamics. J Chem Phys 97:2635–2643.

6. Piana S, Lindorff-Larsen K, Shaw DE (2011) How robust are protein folding simulations with respect to force field parameterization? Biophys J 100(9):L47–L49.

7. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79(2):926–935.

8. Shan Y, Klepeis JL, Eastwood MP, Dror RO, Shaw DE (2005) Gaussian split Ewald: A fast Ewald mesh method for molecular simulation. J Chem Phys 122(5):054101.

9. Frenkel D, Smit B (2002) Understanding molecular simulation, 2nd edition, Academic Press, San Diego.

10. Shaw DE, Dror RO, Salmon JK, Grossman JP, Mackenzie KM, Bank JA, Young C, Deneroff MM, Batson B, Bowers KJ, Chow E, Eastwood MP, Ierardi DJ, Klepeis JL,

Kuskin JS, Larson RH, Lindorff-Larsen K, Maragakis P, Moraes MA, Piana S, Shan Y, Towles B (2009) Millisecond-scale molecular dynamics simulations on Anton. Proceedings of the Conference on High Performance Computing, Networking, Storage and Analysis (SC09). New York: ACM.

11. Piana S, Lindorff-Larsen K, Shaw DE (2013) Atomic-level description of ubiquitin folding. Proc Natl Acad Sci USA 110(15):5915–5920.

12. Heinig M, Frishman D (2004) STRIDE: A web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic Acids Res 32:W500–W502.

13. Wei Y, Thompson J, Floudas CA (2012) CONCORD: A consensus method for protein secondary structure prediction via mixed integer linear optimization. Proc R Soc A 468:831–850.

Supplementary Table and Figures

Number of clusters *K*	Centroid frame of largest cluster (μs)	RMSD of centroid from native conformation (Å)
5	131.42	3.06
10	131.43	2.88
15	129.62	3.89
20	126.49	3.34
25	129.92	3.36
30	129.20	3.78
35	129.60	4.07
40	129.20	3.78
45	129.08	3.68
50	129.20	3.78
55	126.50	3.50
60	129.09	3.82
65	126.50	3.50
70	133.26	2.69
75	129.20	3.78
80	122.40	3.64
85	123.47	3.88
90	129.20	3.78

Table S1. Cluster analysis of frames satisfying all 8 contacts in the simulation of the extended state of Tc684. At each value of the total number of clusters *K* (first column), the table lists the time of occurrence of the centroid frame of the largest cluster (second column) and the RMSD of this centroid frame from the native state of Tc684 (third column).

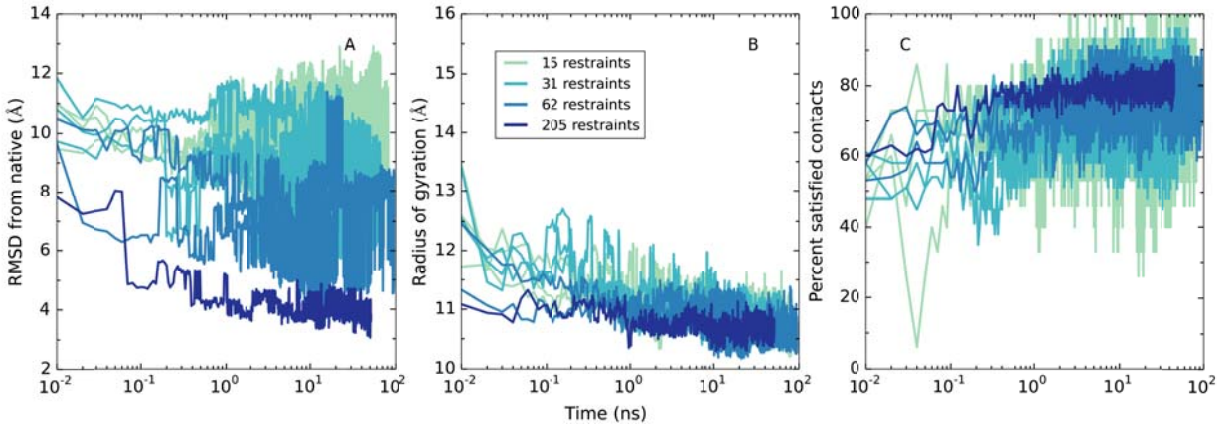

Figure S1. Replica exchange simulations of ubiquitin in vacuo in the presence of distance restraints. These plots display data from the lowest rung (T = 300 K) of the replica exchange ladder for 10 simulations: three separate simulations with 15 randomly chosen distance restraints, three with 31 distance restraints, three with 62 distance restraints, and a single simulation with all 205 distance restraints. The overall trend is that these simulations yield structures that are more native-like (A), more compact (B), and satisfy a larger number of restraints (C) than the starting structures.

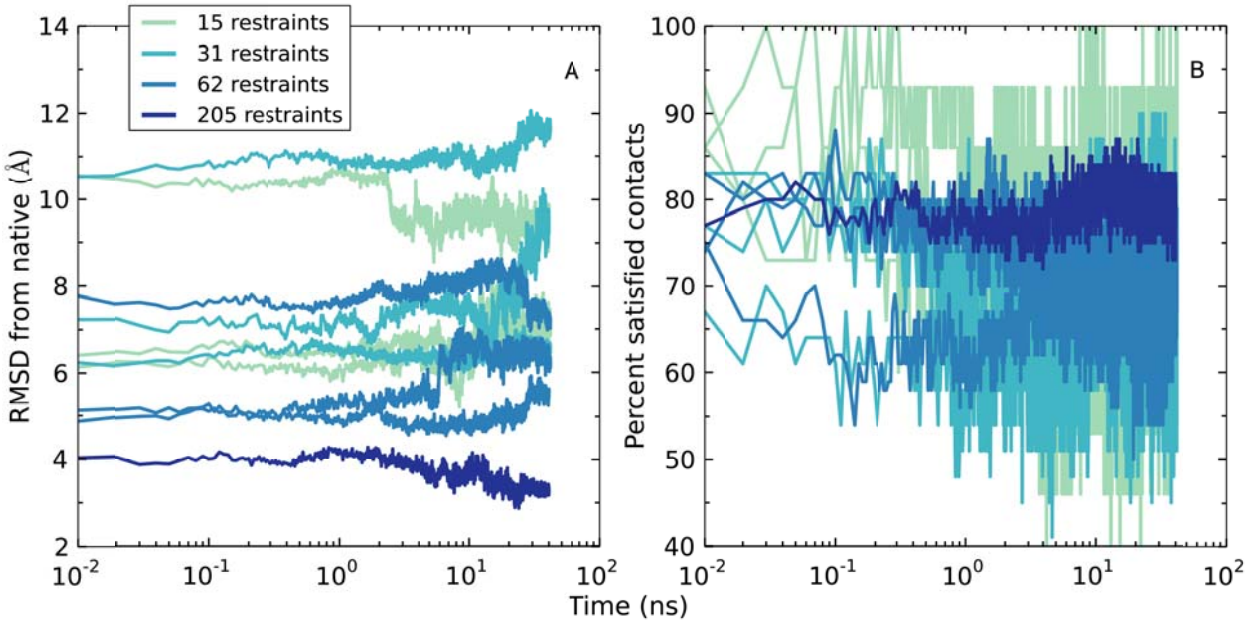

Figure S2. All-atom simulated annealing simulations of solvated ubiquitin in the presence of distance restraints. These simulations start from snapshots taken from the corresponding replica exchange phase, as described in the text. As in Figure S1, ten simulations are represented here. The simulated annealing simulations are carried out in order to relax the structure in the presence of solvent, but no significant trend in either RMSD from the native state (A) or in the number of satisfied contacts (B) is observed over the course of the simulations.

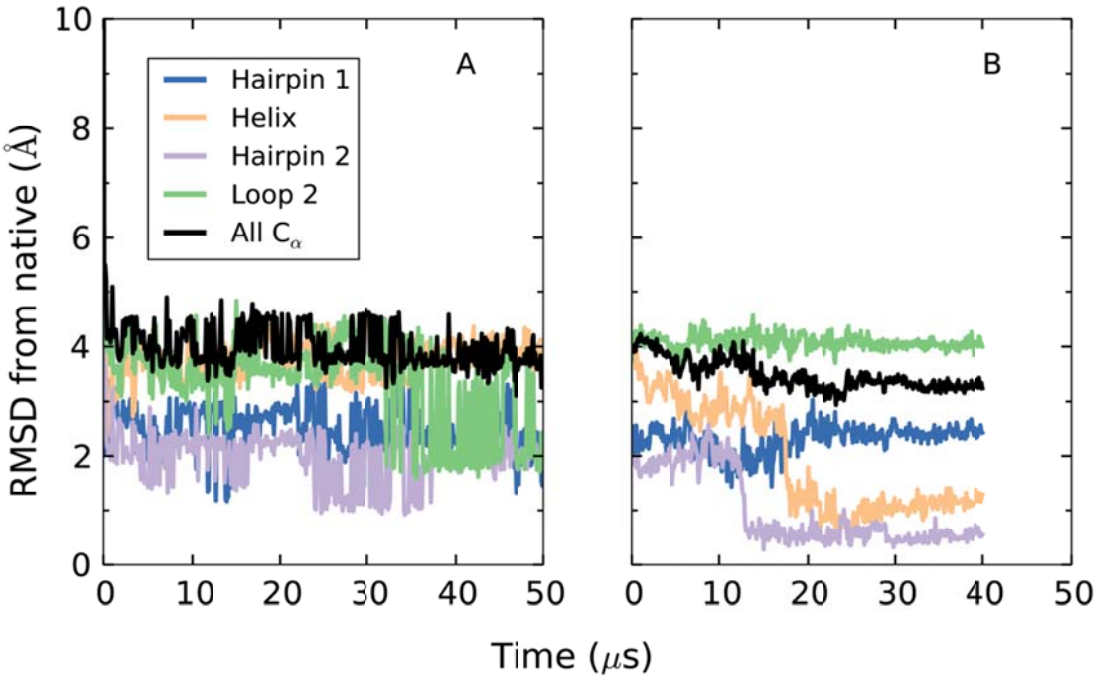

Figure S3. Pre-processing simulations of ubiquitin with all 205 distance restraints present, initially without solvent (A), followed by annealing of the solvated structure (B). Hairpin 2 and the helix are completely formed during the annealing phase.

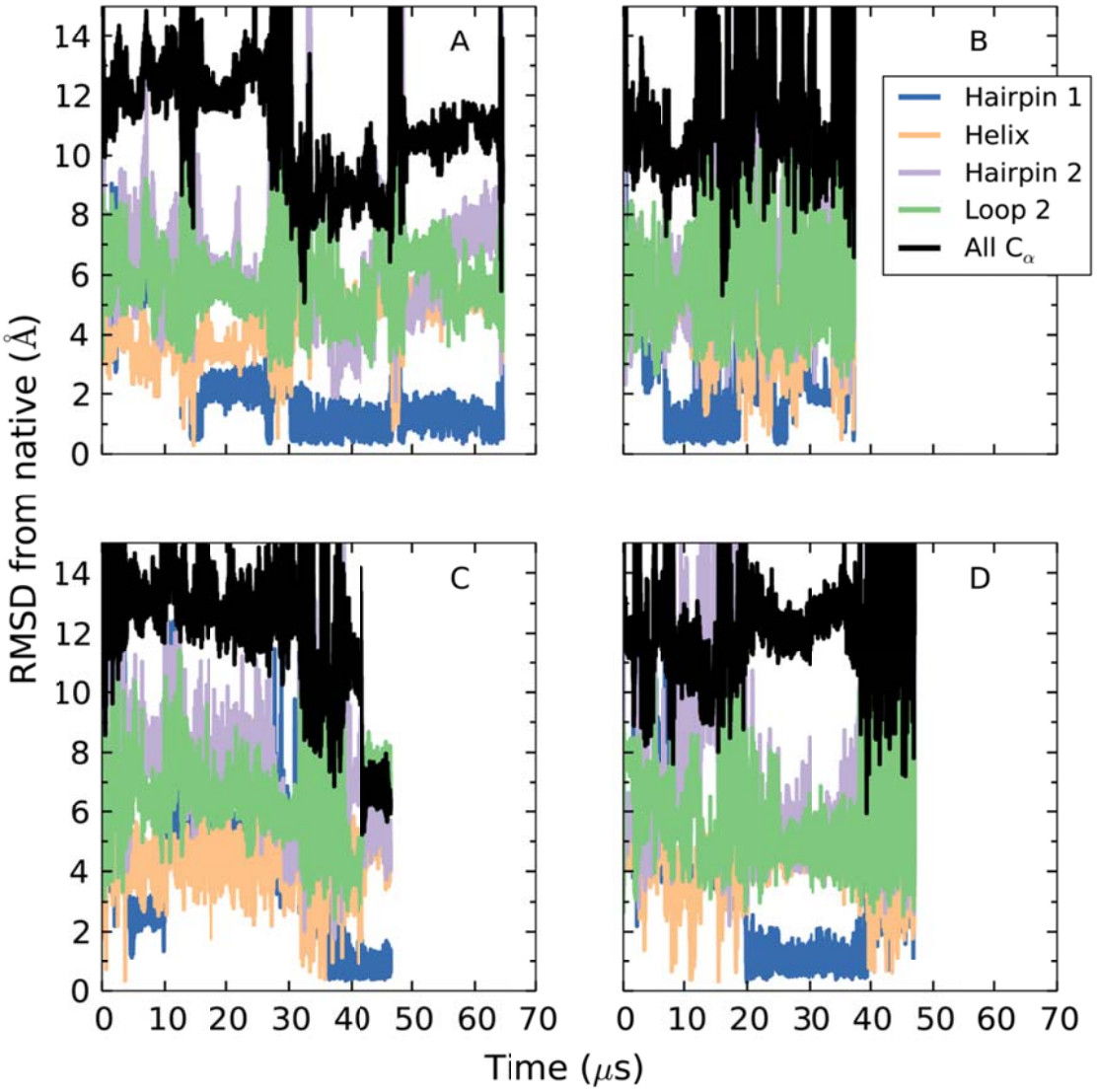

Figure S4. Four all-atom, unrestrained simulated tempering simulations of ubiquitin on Anton. The structure does not converge to a native-like conformation within about 200 μs of total simulation time, although the first hairpin is always formed within the first 40 μs.

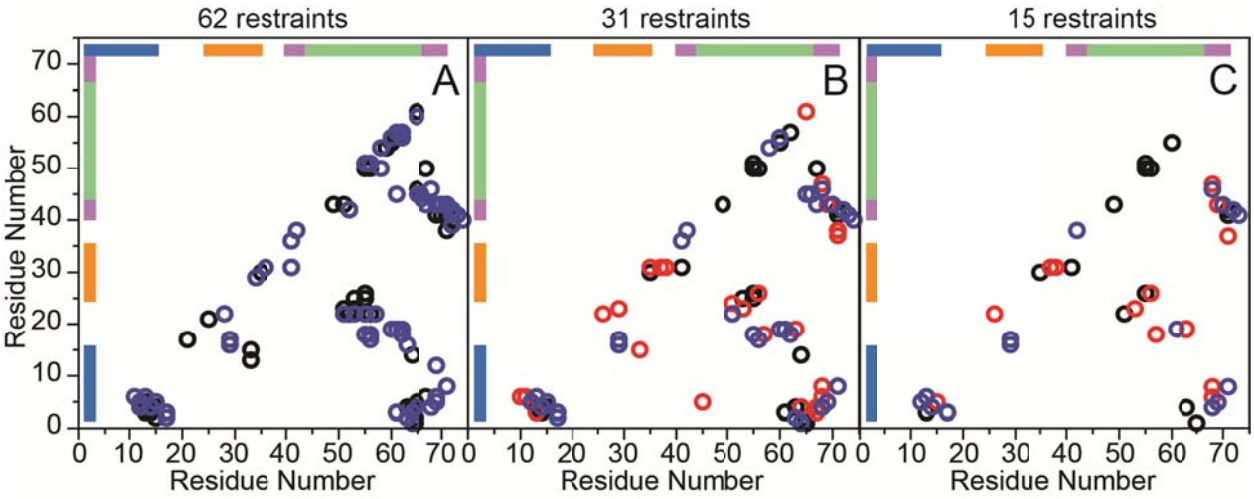

Figure S5. Restraint sets used in the different simulations. (A) 62 restraints; simulation described in Figure 2A (black) and simulations described in Figure 2B–C (blue); (B) 31 restraints; simulation described in Figure 2D (black), simulation described in Figure 2E (red), and simulation described in Figure 2F (blue); (C) 15 restraints; simulation described in Figure 2G (black), simulation described in Figure 2H (red), and simulation described in Figure 2I (blue). The positions of the secondary structure elements of ubiquitin are indicated, colored as in Figure 1.

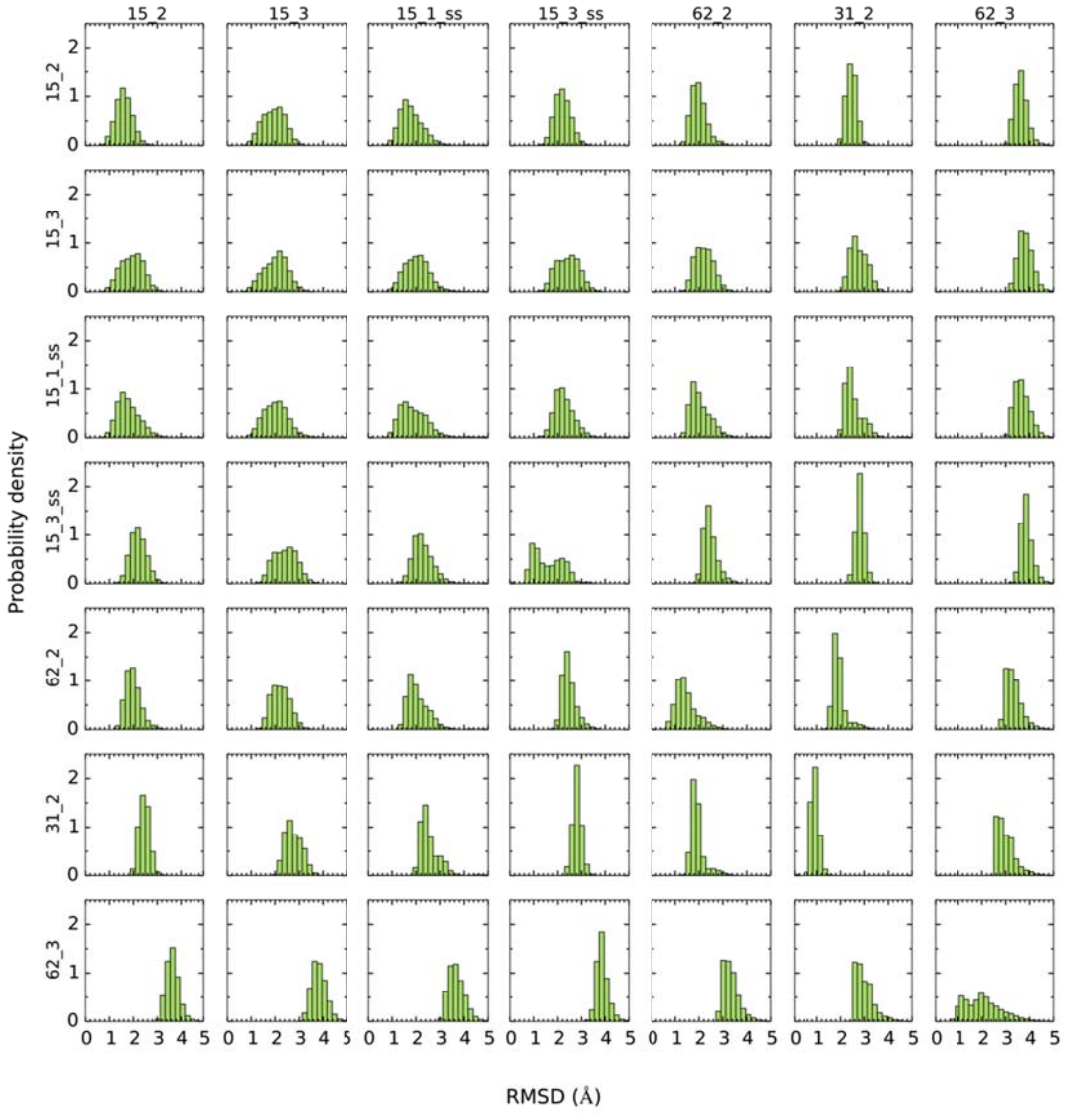

Figure S6. Distributions of all-against-all RMSDs between kinetically trapped ensembles reached in restrained simulations. Each sub-plot is a normalized histogram of all-against-all RMSDs between two kinetically trapped ensembles corresponding to the row and column labels. Row and column labels refer to the simulations. The label '31_2' thus identifies the second simulation with 31 restraints, while '15_3_ss' identifies the third simulation with 15 distance restraints as well as dihedral restraints based on predicted secondary structure. In the RMSD

calculations, snapshots from the following time ranges were chosen from each simulation: 89.5 μs to 105.5 μs for ‘15_2’, 31.7 μs to 49.7 μs for ‘15_3’, 40.1 μs to 50.5 μs for ‘15_1_ss’, 41.9 μs to 44.9 μs for ‘15_3_ss’, 40.1 μs to 49.3 μs for ‘62_2’, 15.0 μs to 25.0 μs for ‘31_2’, and 8.9 μs to 40.1 μs for ‘62_3’.

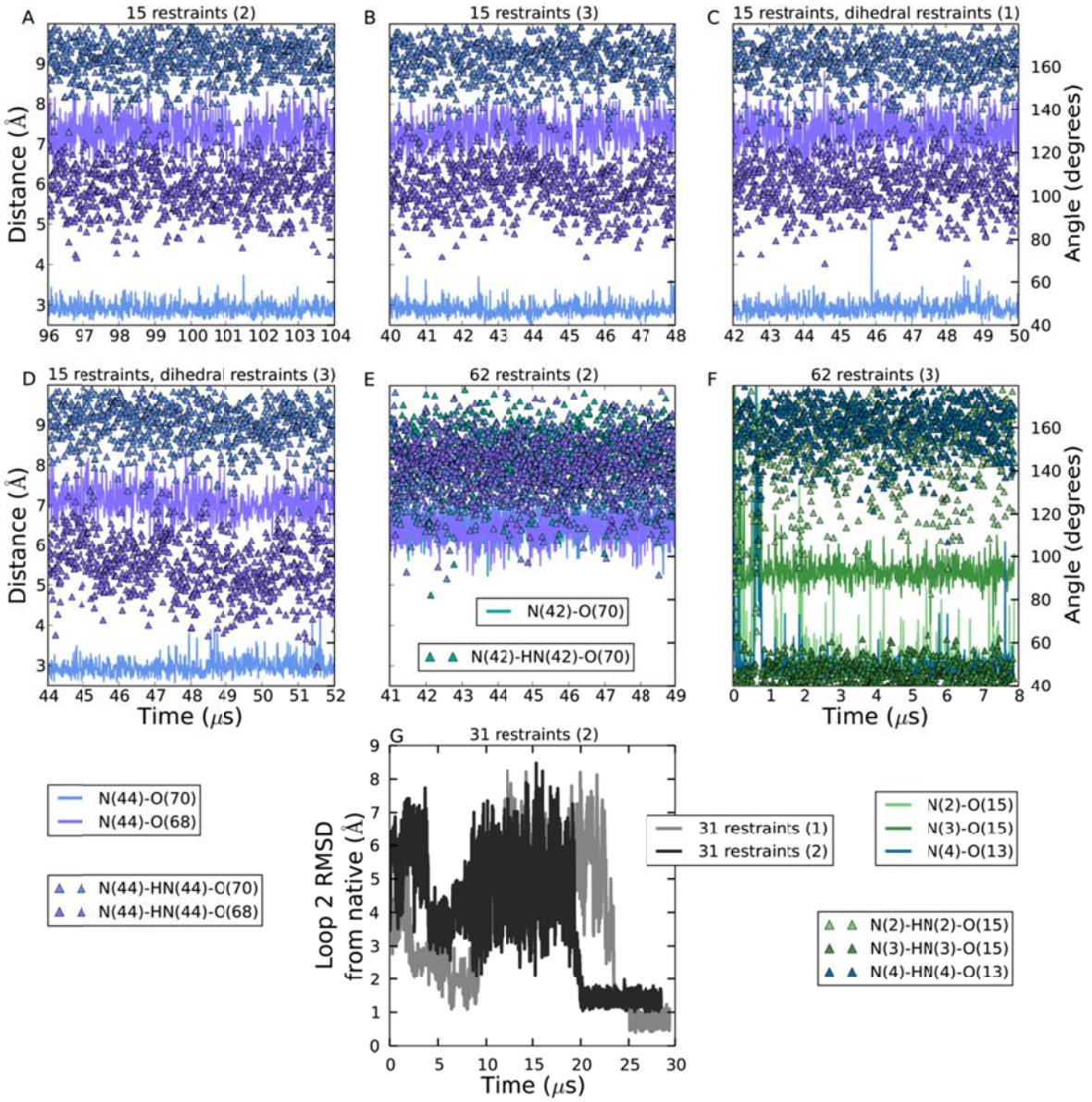

Figure S7. Hydrogen bond angles and distances (A–F) and conformational changes in loop 2 (G) in kinetically trapped ensembles. Native ubiquitin has hydrogen bonds between residues 3 and 15 (β hairpin 1) and between residues 44 and 68 and residues 42 and 70 (β hairpin 2). The native hydrogen bonds in hairpin 2 are broken in four kinetically trapped ensembles, and hydrogen bonds between residues 44 and 70 are formed, signaling a register shift in hairpin 2 (A–D). In E, all native hydrogen bonds in hairpin 2 are broken and no new hydrogen bonds are formed. In F, there is a register shift in hairpin 1 signaled by the breaking of native hydrogen

bonds and the formation of hydrogen bonds between residues 2 and 15 and residues 4 and 13. In the '31_2' simulation (G), the kinetically trapped state is characterized by the modified conformation of loop 2 relative to that in the native state (reached by the '31_1' simulation).

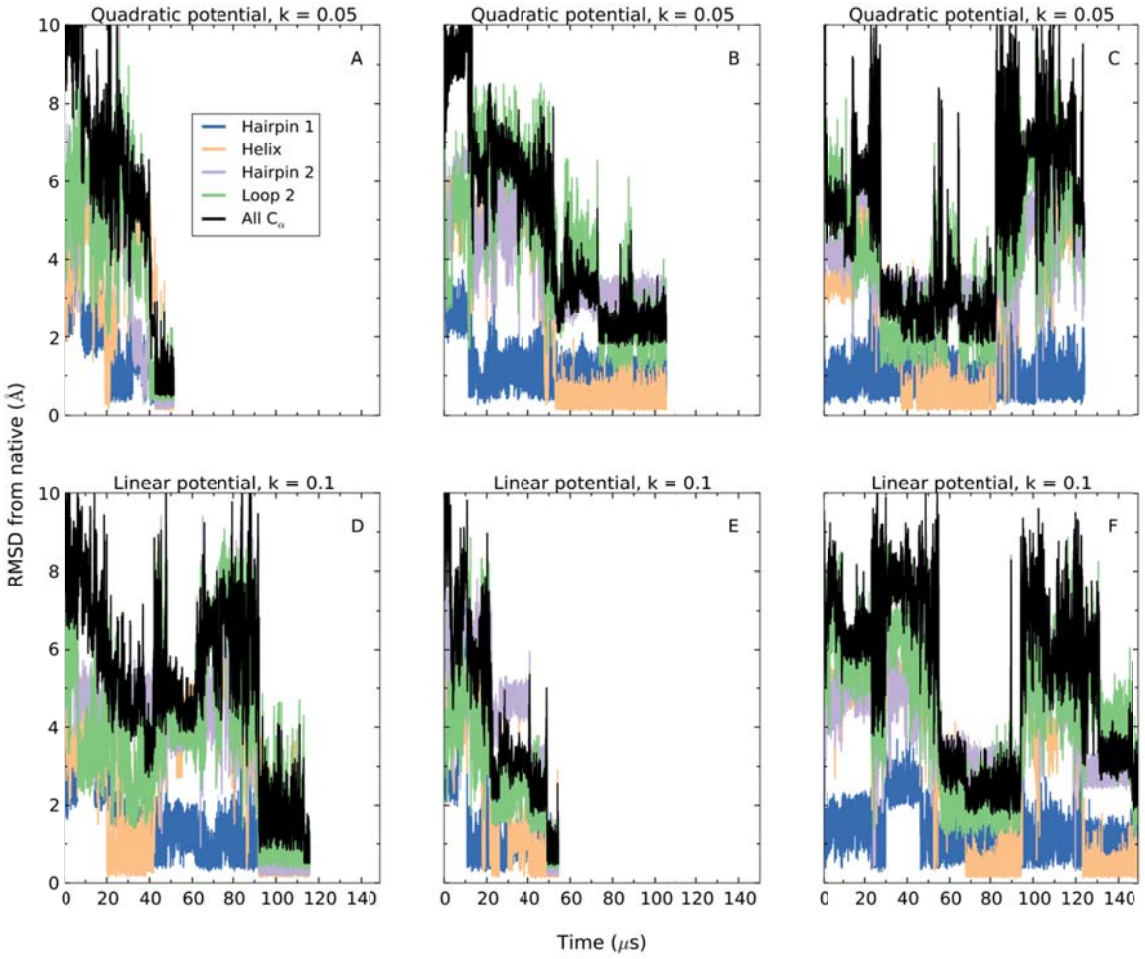

Figure S8. Comparison of flat-bottomed harmonic (A–C) with flat-bottomed linear (D–F) potentials in restrained simulated tempering simulations of ubiquitin starting from extended conformations. All simulations have 15 distance restraints, and vertically aligned plots correspond to the same set of distance restraints. A and D, for example, thus have the same set of 15 distance restraints. The simulations converge to the native conformation in two out of three cases with a linear potential as opposed to only one case with a quadratic potential, although there appears to be no significant gain in the rate of convergence.

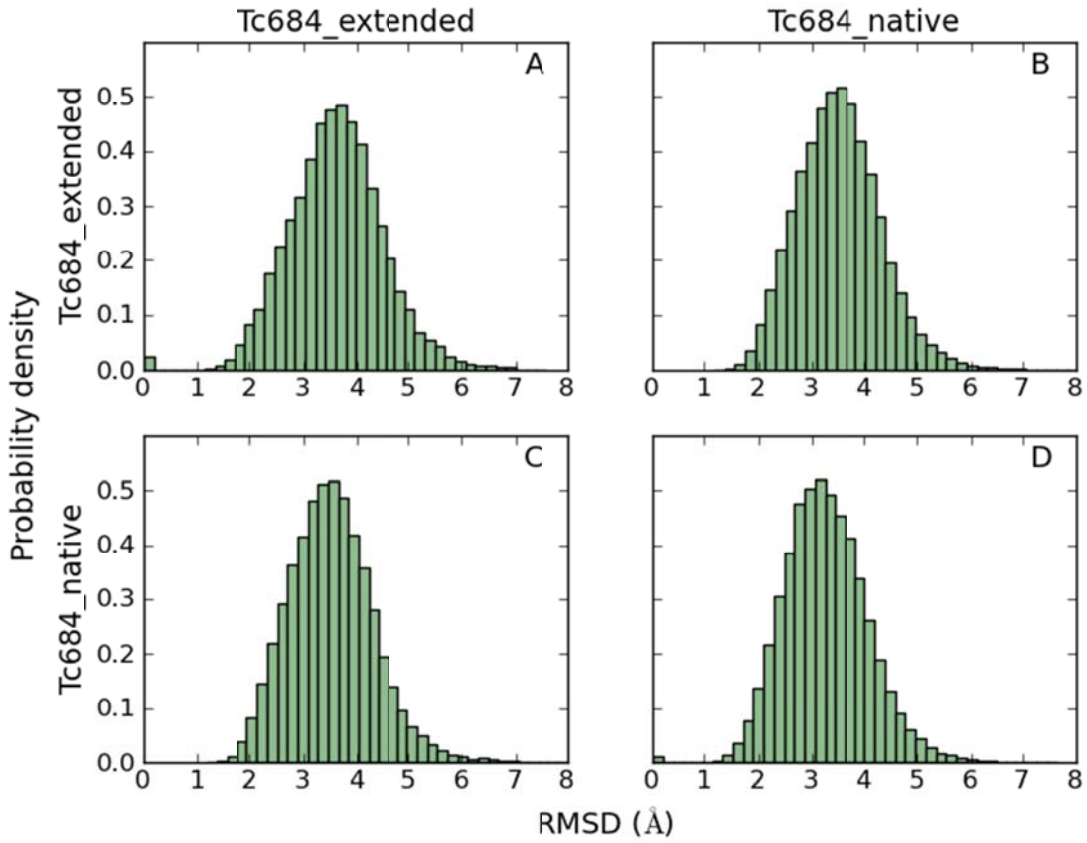

Figure S9. Normalized histograms of all-against-all RMSDs obtained using a set of frames from the native state simulation of CASP target Tc684 and a set of frames taken between 135 μs and 145 μs of the extended-state simulation of the same target with both distance and dihedral restraints (the simulation of Figure 5D in the main text). Plots (A) and (D) correspond to RMSD values between frames within the extended and native state simulations, respectively, while the identical plots (B) and (C) correspond to RMSD values between frames in native and extended state simulations. The similarity of these histograms shows that the extended-state simulation between 135 μs and 145 μs has essentially reached the native ensemble.