Supporting Information

Smith et al. 10.1073/pnas.1519609113

SI Methods

Sample Preparation. Perdeuterated, ^{15}N-labeled WT and mutant ubiquitin was expressed in *Escherichia coli* adapted to 100% D_2O Toronto minimal medium supplemented with D_7-glucose as a carbon source and ^{15}N-NH_4Cl as a nitrogen source. The ubiquitin mutants E24A and E53A were generated by PCR-based, site-directed mutagenesis using the QuikChange II Kit (Agilent) following the instructions of the supplier. The catalytic core of USP2 (amino acids 259–605) was expressed and purified following published protocols (32).

NMR. The ^{15}N $R_{1\rho}$ measurements were conducted using uniformly ^{15}N-labeled ubiquitin in 90% (vol/vol) H_2O/10% (vol/vol) D_2O. The procedure used here followed previously published methods (12, 33). Field strengths were varied from 1,000 to 6,000 Hz. Rates were determined using a two-point sampling scheme in which one reference experiment was recorded without any spin-lock period and another with a spin-lock applied for 120 ms. The errors in rates were propagated from noise in the spectra. In this study, data for T14, L43, and F45 were acquired, adding to the six previously measured dispersion curves (Fig. S1*A*).

The $^1H^N$ $R_{1\rho}$ was collected using a U-[^{15}N]–labeled sample of perdeuterated ubiquitin in 90% (vol/vol) H_2O/10% (vol/vol) D_2O. For the measurement of ^{1}H $R_{1\rho}$, the experiments used follow previous methods (34). Spin-lock frequencies were varied from 1,000–10,000 Hz (277 K for WT) or 27,000 Hz (other temperatures and mutants), and were calibrated by measuring ^{1}H 90° pulse lengths at their corresponding power levels. Field strengths and offsets were chosen such that tilt angles of ∼35° were used for all points to minimize the nuclear Overhauser effect (NOE) and rotating frame nuclear Overhauser effects (ROE) that can lead to pseudodispersion profiles (34). The experiments were carried out in an interleaved fashion, where the used delay, field strength, and offset were randomly varied. With the current experimental parameters, the overall change in the temperature was less than 1 K. Relaxation rates were determined using a three-point sampling scheme with spin-lock relaxation delays of 5, 65, and 125 ms. Rate errors were estimated using residuals from the three-point fits. Mean $R_{1\rho}$ and ω_{eff} (effective radio frequency field) values were determined using equations 5 and 6, respectively, from the study by Eichmüller and Skrynnikov (34). For each experiment, 56–188 (indirect dimension) and 1,024 (direct dimension) complex points were acquired. Four transients for each increment were collected with recycle delays of 2.3 s (277 K for WT) or 2 s (other temperatures and mutants), yielding a total experiment time of 17–58 min per data point. For E24 and G53 RD curves acquired at 308 K, the sweep width was increased from 23.5 ppm (used in other experiments) to 49.3 ppm, and a six-point sampling scheme was used with spin-lock relaxation delays of 5, 14, 23, 32, 41, and 50 ms. All $^1H^N$ experiments were conducted on a Bruker spectrometer operating at a ^{1}H frequency of 600 MHz.

The $^1H^N$ Carr–Purcell–Meiboom–Gill (CPMG) measurements were conducted on a U-[^{15}N] and U-[^{2}H] selectively ^{13}C-labeled (CHD_2) sample of perdeuterated ubiquitin in 90% (vol/vol) H_2O/10% (vol/vol) D_2O. The sample was loaded in nine capillaries placed into a 5-mm sample tube similar to a previously published method (14). Supercooled CPMG experiments were acquired at 262 K using a relaxation compensated approach with the CPMG period following t_1 evolution (35) and phase cycling of the refocusing pulses (36, 37). The CPMG frequency (ν_{CPMG}) was varied from 211 to 2,526 Hz during a constant time period (T_{CP}) of 19 ms. For each experiment, 128 (indirect dimension) and 1,024 (direct dimension) complex points were acquired with a recycle delay of 1 s. For the CPMG experiments, 60 transients were collected for a total experiment time of 315 min per ν_{CPMG} point. For the reference experiment without the CPMG block, eight transients were collected for a total experiment time of 45 min. Rate errors were propagated from spectral noise. Effective relaxation rates ($R_{2,\mathrm{eff}}$) were fit assuming fast exchange using the following formula:

$$R_{2,\mathrm{eff}} = R_2 + \Phi_{ex}\tau_{ex}\left(1 - 4\nu_{CPMG}\tau_{ex}\tanh\frac{1}{4\nu_{CPMG}\tau_{ex}}\right).$$

NMR data processing and peak quantification were done with NMRPipe. Peak intensities for E24 and G53 at 308 K were calculated by taking the maximum spectral intensity in a region around the peak positions. $R_{1\rho}$ parameter fitting, error estimation, and determination of significant amounts of dispersion were done as previously described (13). CPMG data were processed in the same manner.

Optimizing a Single Collective Mode to Explain the RD Data. We also tested a third step in which the resulting weights were further refined using Nelder–Mead optimization. Although this technique produced higher ROC curve areas for the training data, it did not produce better ROC curves using the cross-validation procedure described in *Methods*, *Cross-Validation of the Optimization Procedure*. Therefore, we did not use this procedure in a final determination of optimized weights. Using the same cross-validation scheme, we determined that the optimal number of PCA modes to include in the optimization was 20. The top 20 eigenvectors covered 78% of the variance and 37% of the SD (i.e., atomic displacement) in the underlying coordinates.

Clustering RD Fit MD Mode into Semirigid Bodies. To identify a set of semirigid bodies in the RD fit MD mode, residues were first clustered into contiguous segments along the amino acid sequence and then clustered into discontinuous groups of segments. Backbone rmsds between the minimum and maximum projections were used for clustering. For contiguous segment clustering, residues were initially put into single-residue segments. Using a greedy algorithm, the pairs of adjacent segments having the lowest combined rmsd were successively merged. A cutoff of 0.35 Å was used to create 11 contiguous segments for group clustering. Clustering of these segments into discontinuous groups was done with the same greedy algorithm, except that merged groups were no longer required to be adjacent in sequence. A cutoff of 0.7 Å was used to create four rigid bodies.

Analysis of Peptide Flipping in MD Trajectories. For every snapshot of the 100-ns AMBER ff99SB MD trajectories (10) or the 1-ms CHARMM22* MD trajectory (22), we calculated the ψ-backbone dihedral angle of D52 and the ϕ-backbone dihedral angle of G53. To ensure that structural transitions did not wrap around this periodic ψ/ϕ-space, the angles were normalized in the following manner. First, both angles were mapped onto the range 0–360°. Second, for any ψ_i/ϕ_i-pair whose sum was greater than 350°, ψ_i was set to $\psi_i - 360$. Under this normalization, $\psi_i - \phi_i$ is centered around 80° in the NH-in conformation and −320° in the NH-out conformation.

For the CHARMM22* trajectory, the alternate states were identified by inspection of the rmsds of residues 51–53 (alternate state 1) and residues 31–41 (alternate state 2). Q-factors were

determined from a combined set of backbone NH, backbone NC, and side-chain residual dipolar couplings (RDCs), using a single alignment tensor for all RDCs in a given alignment medium. To do so, a five-element B vector was first calculated (42) for each internuclear vector in every snapshot in the trajectory. From these vectors, average B vectors for each state were then calculated. The landscape of Q-factors was determined by varying the weights applied to the average B vectors for each state.

PLS FMA. To enable Rosetta structural modeling of the entire core ubiquitin structure (residues 1–70), we first calculated projection values for the 209 non–USP-bound ubiquitin conformations onto the peptide fit PDB mode. We then used these projection values (instead of the previous 0 or 1 assignments) to train a new PLS model using the backbone N, CA, and C atoms of residues 1–70. The use of 40 PLS components resulted in near-exact reproduction of the peptide fit PDB mode projection values. For each of the 217 crystallographic conformations, we used this PLS model to generate 21 synthetic conformations evenly interpolated from the minimum (NH-in) to the maximum (NH-out) projection value.

USP2 Inhibition Assays. USP2 inhibition assays were performed similar to previously described assays (32), in which WT and modified ubiquitin constructs inhibit cleavage of 7-amido-4-methylcoumarin (AMC) C-terminally linked to WT ubiquitin. The final reaction buffer included 100 mM NaCl, 50 mM Hepes (pH 7.6), 25 μM $ZnCl_2$, 5 mM DTT, and 8% DMSO. Thirty-microliter reactions were performed at 25 °C in low-volume, 384-well, black with clear bottom NBS microplates (Corning). Initial rates of fluorescence increase were monitored with a PHERAstar FS microplate reader (BMG Labtech) using a 340-nm/470-nm filter pair. The USP2 concentration was determined from its absorbance at 280 nm and kept fixed at 1.5 nM for all reactions. For K_m determination, concentrations of ubiquitin-AMC (Boston Biochem) were log-spaced from 0.002 to 9.3 μM (12 points in duplicate), without the presence of inhibitor. For determination of K_i values, the concentrations of WT, E24A, and G53A ubiquitin were log-spaced from 0.1 to 1,000 μM (13 points in triplicate), with a fixed concentration of 2 μM ubiquitin-AMC substrate. Inhibitor concentrations were determined by mass. Kinetic parameters were calculated by simultaneously fitting the maximum velocity (V_{max}), Michaelis–Menten constant (K_m), and inhibition constants ($K_{i,WT}$, $K_{i,E24A}$, and $K_{i,G53A}$), with the following equation:

$$V = \frac{V_{max}[S]}{K_m\left(1 + [I_{WT}]/K_{i,WT} + [I_{E24A}]/K_{i,E24A} + [I_{G53A}]/K_{i,G53A}\right) + [S]}.$$

Input data included the initial reaction velocities (V), substrate concentrations ($[S]$), and inhibitor concentrations ($[I_{WT}]$, $[I_{E24A}]$, and $[I_{G53A}]$, with no more than one inhibitor per reaction). After this global fitting, errors in the inhibition constants were analytically determined by individually fitting each K_i with values of V_{max} and K_m taken from the global fit. For pairs of sample data (WT/E24A or WT/G53A), ANOVA F test P values were determined by comparing a fit to the global equation with one where the K_i was assumed to be the same for both samples. For these fits, V_{max} and K_m were also taken from the initial global fit.

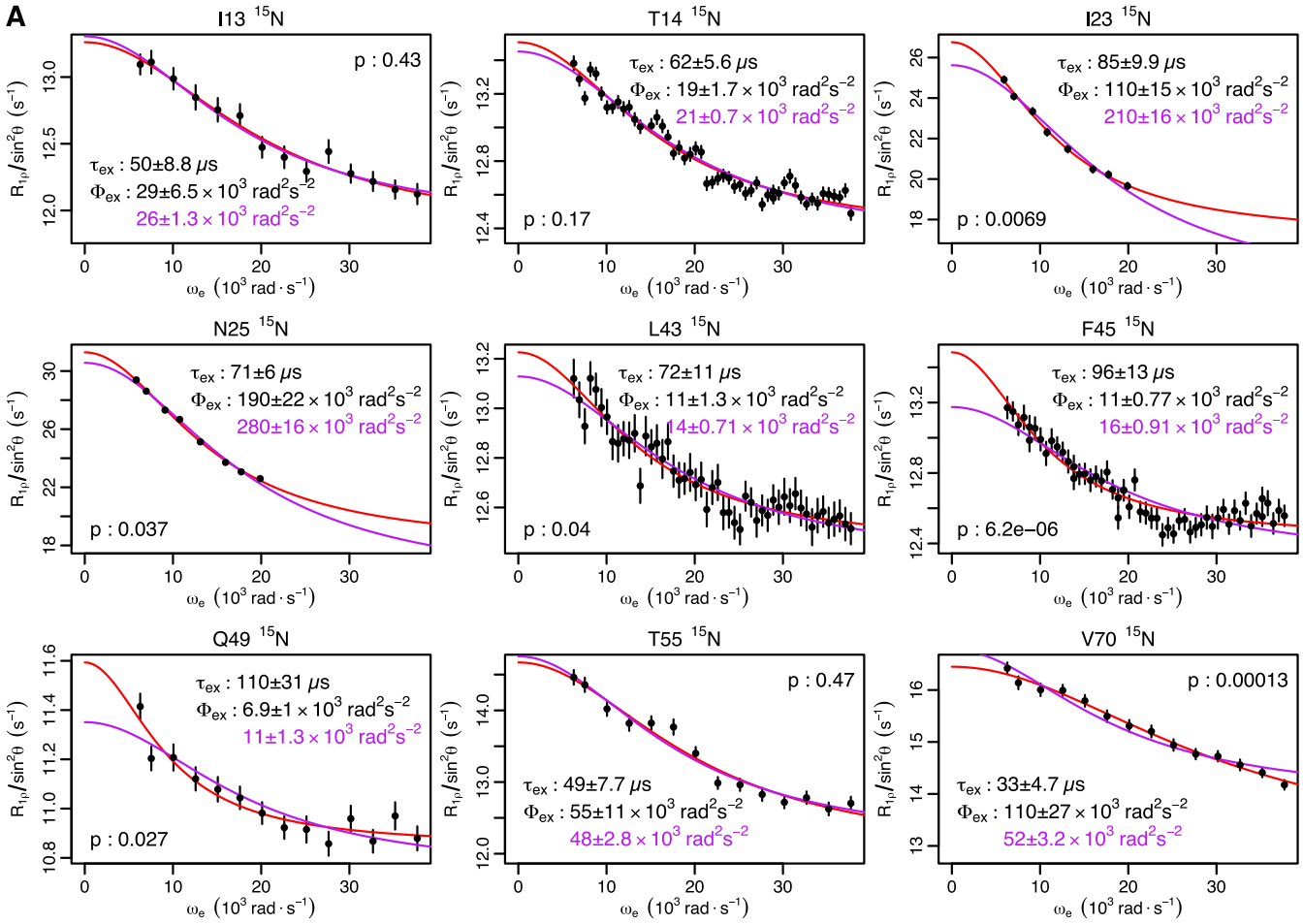


Fig. S1. (Continued)

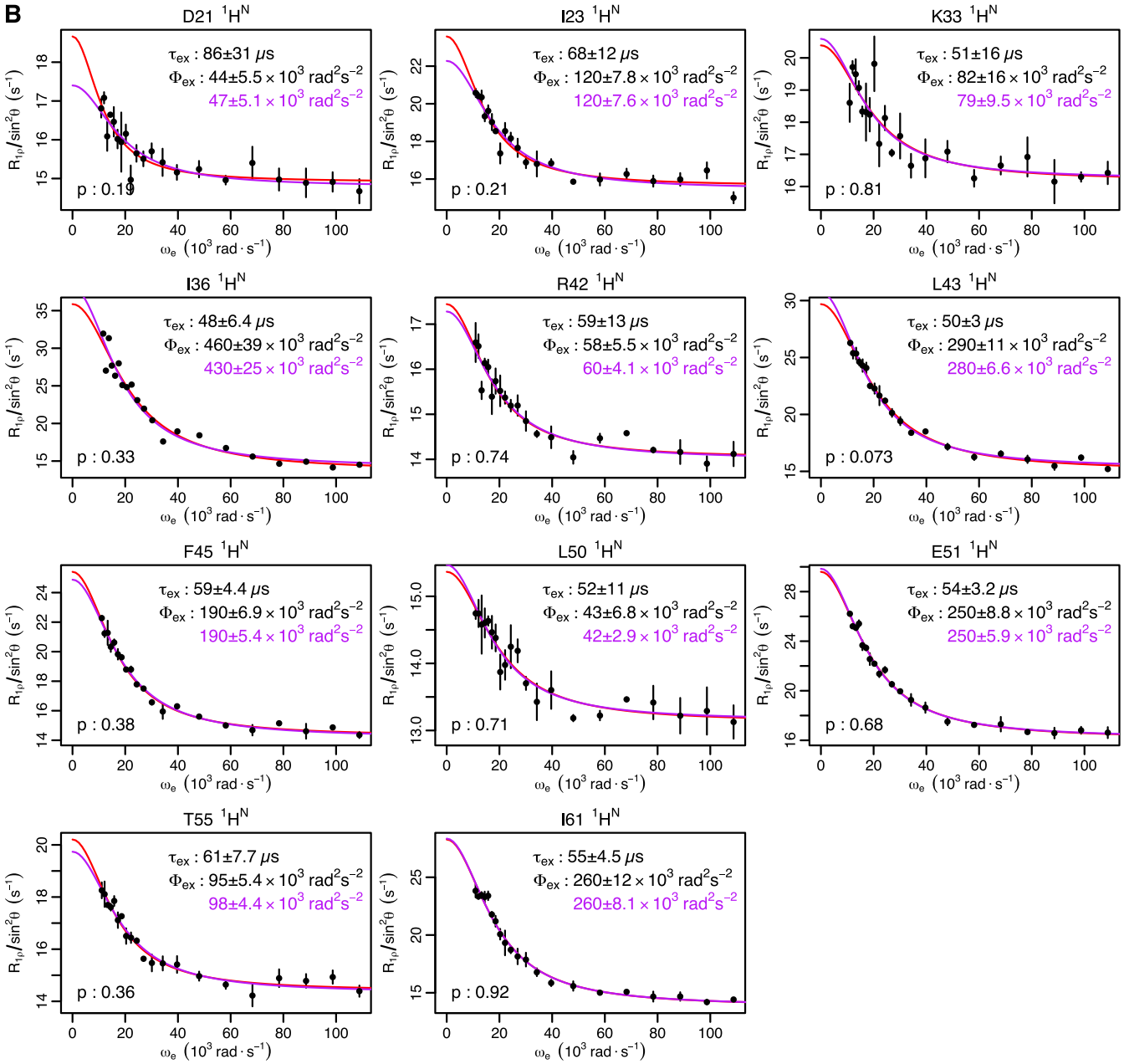


Fig. S1. (Continued)

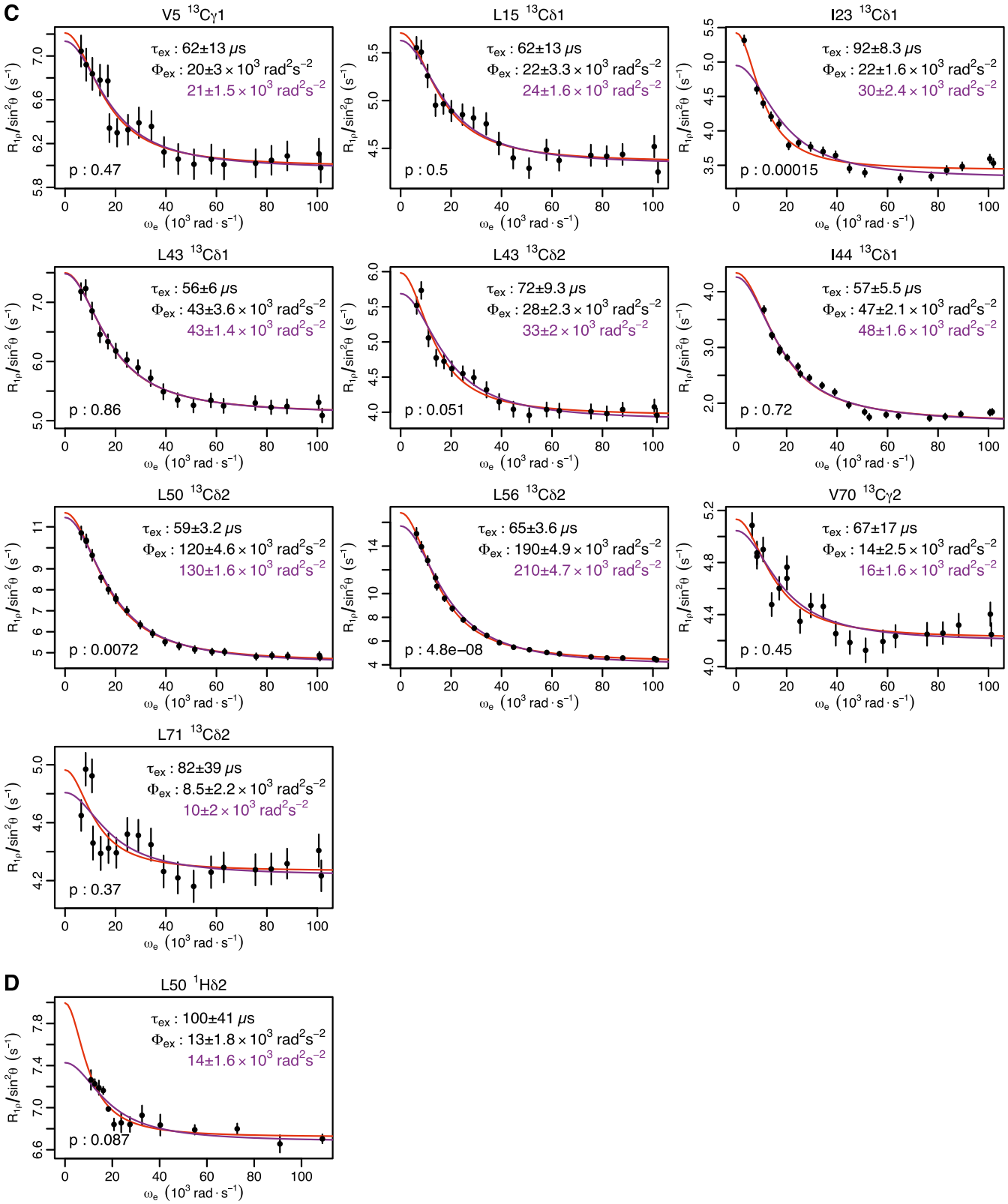


Fig. S1. Ubiquitin $R_{1\rho}$ data at 277 K. Individual fits are shown in red, with the parameters shown in black. Global fits with a single τ_{ex} (55 μs) are shown, along with the corresponding Φ_{ex} in purple. *F* test *P* values between fits are also shown. (*A*) Backbone ^{15}N: T14, L43, and F45 are from this study. I23 and N25 are from previous work (14). I13, Q49, T55, and V70 are from previous work (12). (*B*) Backbone ^{1}H^{N} from this study. (*C*) Methyl ^{13}C from previous work (13). (*D*) Methyl ^{1}H from previous work (13).

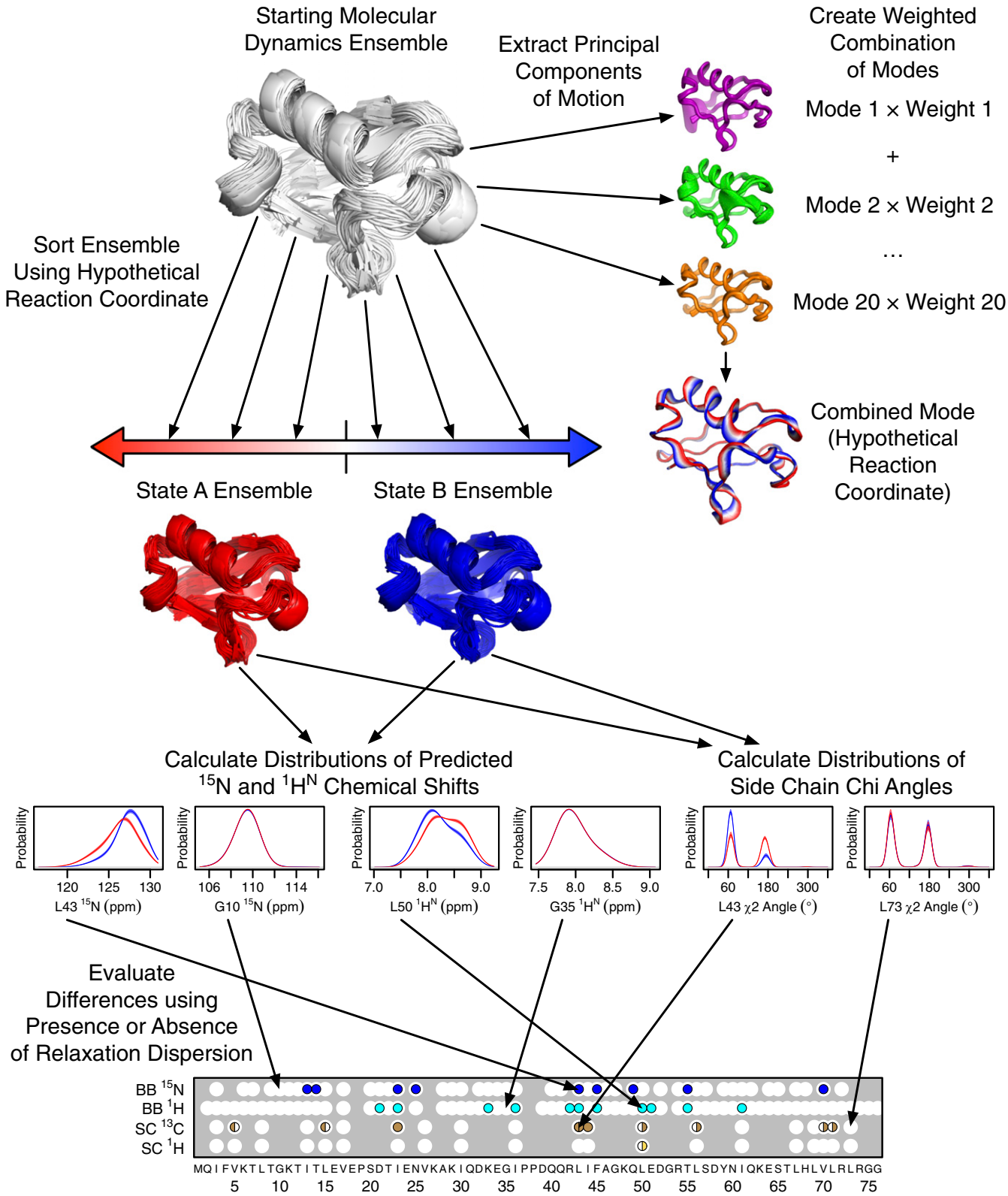


Fig. S2. Scheme for generating RD fit MD mode. In this study, a MD ensemble was used, but in principle, any ensemble with sufficiently dense sampling could be used. Error bars in the distributions of predicted chemical shifts or χ-angles are indicated by shaded regions around the lines. The differences between the distributions (shown in Fig. S4) were calculated by determining the area between the error bars of the two distributions.

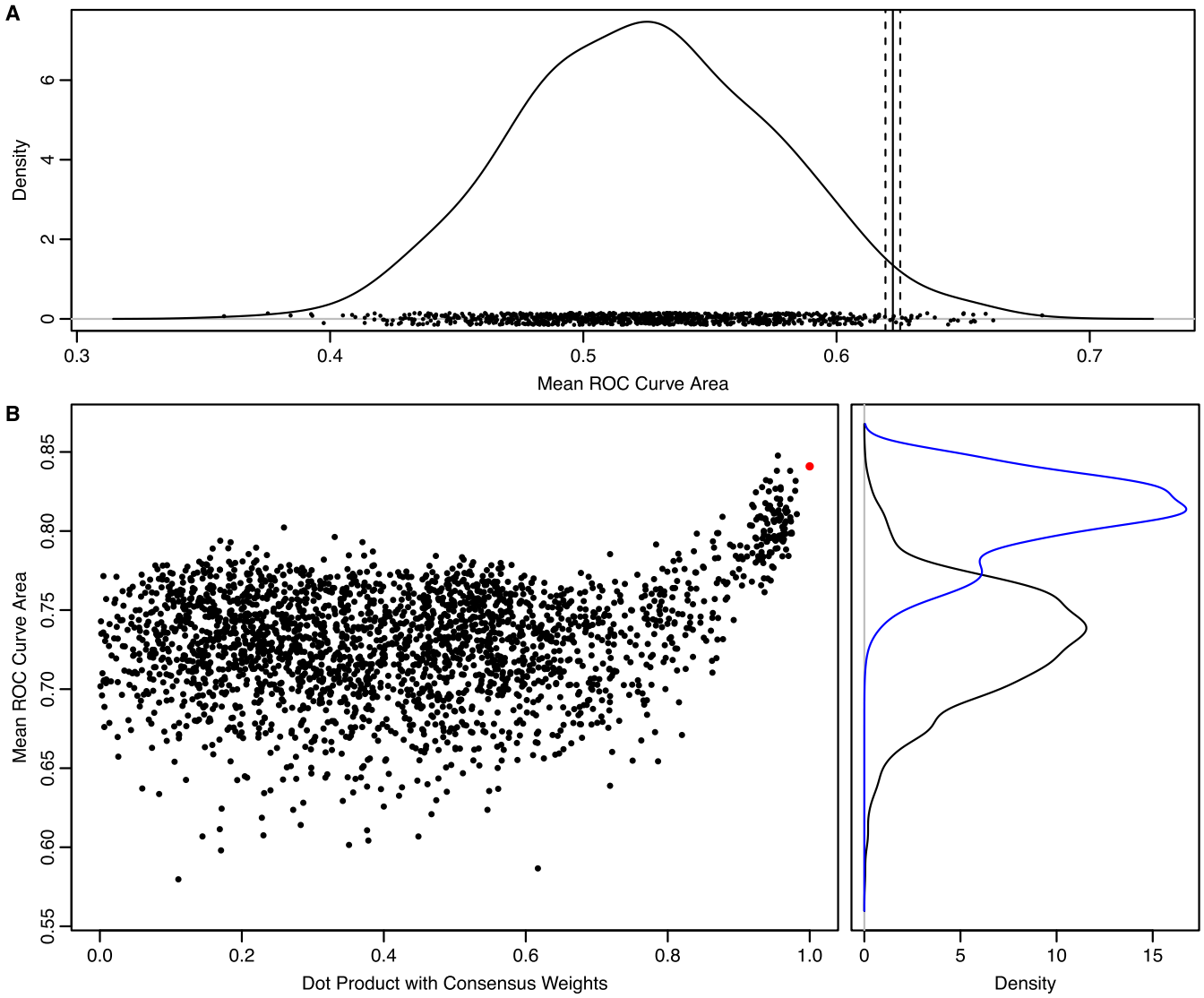


Fig. S3. Cross-validation of RD fit MD mode fitting. (*A*) Distribution of cross-validated mean ROC curve areas for shuffled experimental RD data. The distribution is derived using kernel density estimation with a Gaussian smoothing kernel having an SD equal to the average SE of the underlying data points (0.015). The underlying data points are shown at the base of the plot. The mean ROC curve area for unshuffled experimental data is shown as a solid vertical line, along with the associated SE (dashed vertical lines). The probability of observing a shuffled value larger than the unshuffled value is 0.027. (*B*) Rescoring of the different weight vectors derived from the cross-validation runs using the full set of experimental data (*Left*, black points) produces a multimodal distribution with a secondary peak having a higher mean ROC curve area (*Right*, black curve). Consensus weights were determined using PCA with a weighted covariance matrix strongly biased toward points in the secondary peak. The distribution of mean ROC curve areas using those weights is shown in blue. The consensus weights produce a higher mean ROC curve area (red point) than all but one of the cross-validation weights. The single best set of cross-validation weights is very similar to the consensus weights. The differences between the consensus weights and the different cross-validation weights are shown using the absolute value of the dot product of the respective vectors. This plot shows strong funnel-like behavior (upside down because higher values are better), suggesting that the consensus weights are close to the global minimum of the optimization landscape.

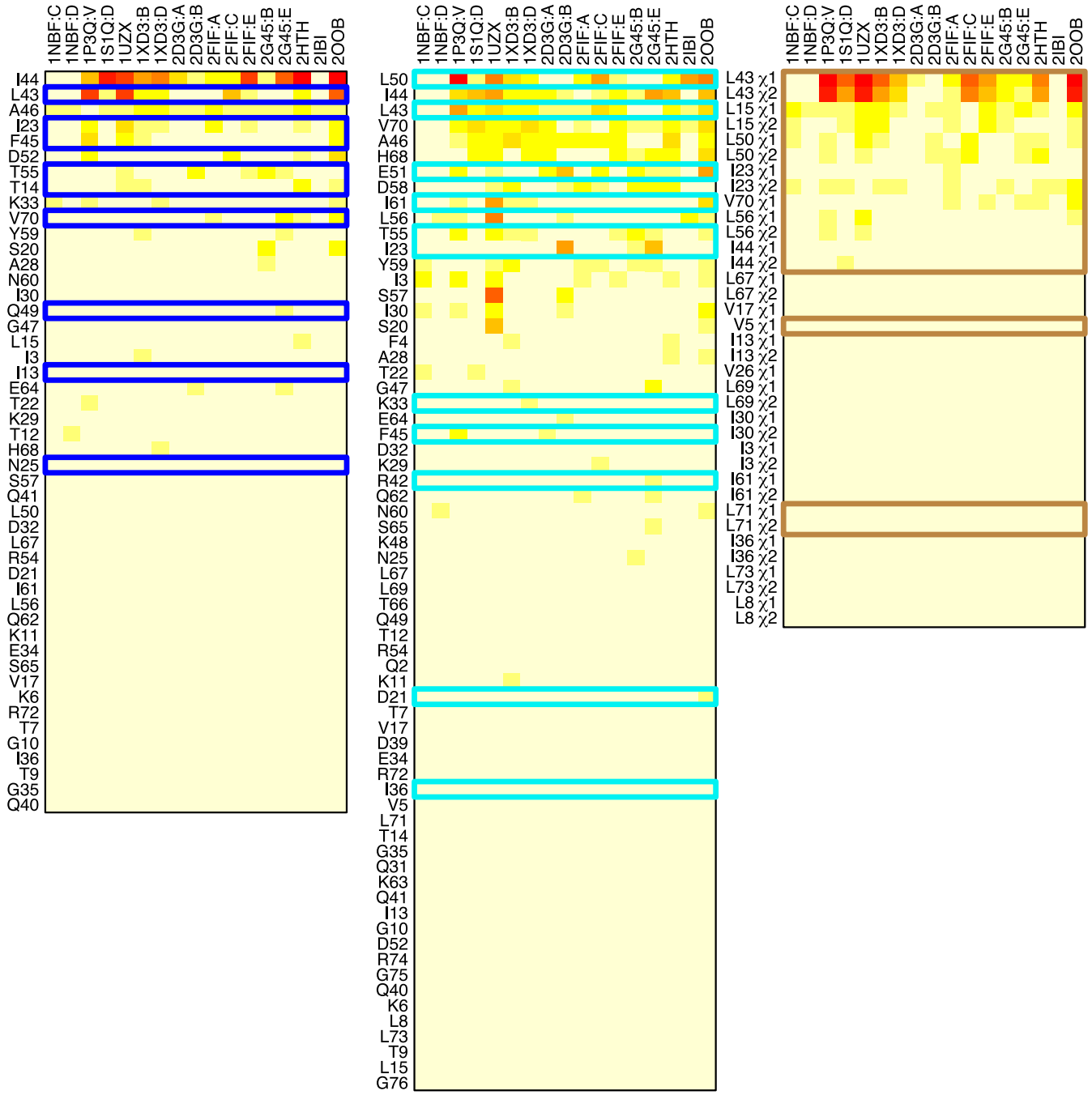


Fig. S4. Chemical shift and χ-angle differences predicted using the RD fit MD mode. Distribution differences between MD simulation subensembles (red and blue in Fig. S2) are shown for predicted amide ^{15}N chemical shifts (*Left*), predicted amide ^{1}H^{N} chemical shifts (*Middle*), and side-chain χ-angles (*Right*). Red indicates the greatest difference, and pale yellow indicates the least difference. Columns correspond to MD simulations started from the indicated PDB structure and chain. Rows are sorted by the mean difference for each residue. Residues with significant experimentally observed RD (i.e., those residues for which there should be larger differences in predicted chemical shifts or χ-angles) are outlined with colored rectangles. If the predictions were perfect, there would be a single colored rectangle at the top of each grid, and it would be possible to set a threshold having a 100% true-positive rate without any false-positive rate.

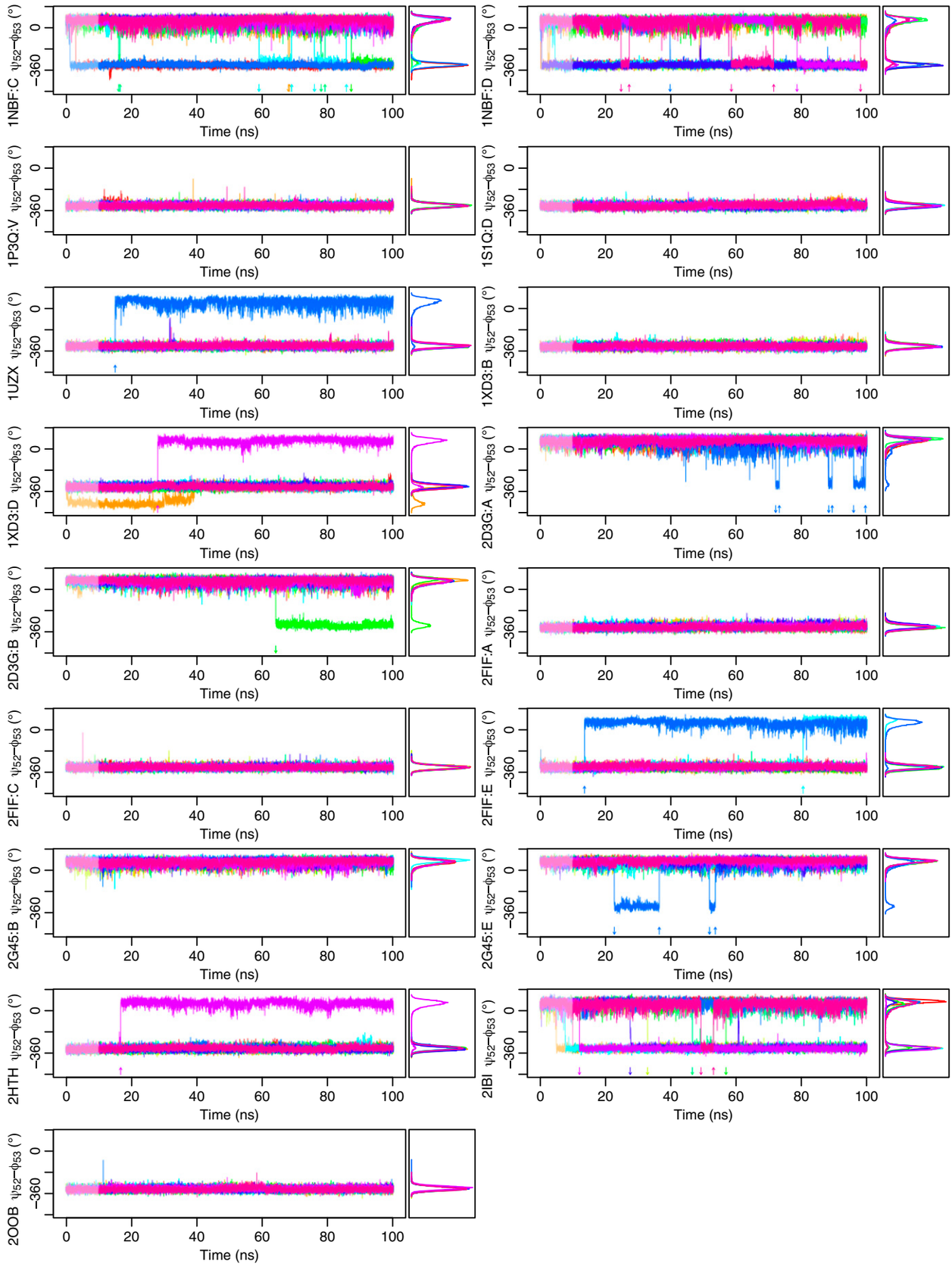


Fig. S5. Peptide flipping is observed in 100-ns AMBER ff99SB trajectories. In the 100-ns simulations, the peptide bond between residues D52 and G53 rarely flips (indicated by arrows). The orientation is quantified by the difference between the ψ_{52} and ϕ_{53} dihedral angles. When the amide proton is pointed toward the α-helix and can hydrogen-bond with E24, the difference is ∼80°. When the amide proton is pointed out into solvent, the difference is approximately −320°. Each of the 10 replicate simulations is shown with a different color. Lightened colors are used to show the first 10 ns of each simulation, which were discarded for calculation of the χ-angle and chemical shift distributions. Despite the peptide bond rotation being one of the slowest processes in the simulations, its rate is ∼100-fold faster than is observed experimentally. This difference suggests that either the barrier in the simulation is too small (perhaps on account of the torsional potential) or that the individual states are understabilized (due to hydrogen bonding or other features) on the order of 1–3 kcal/mol.

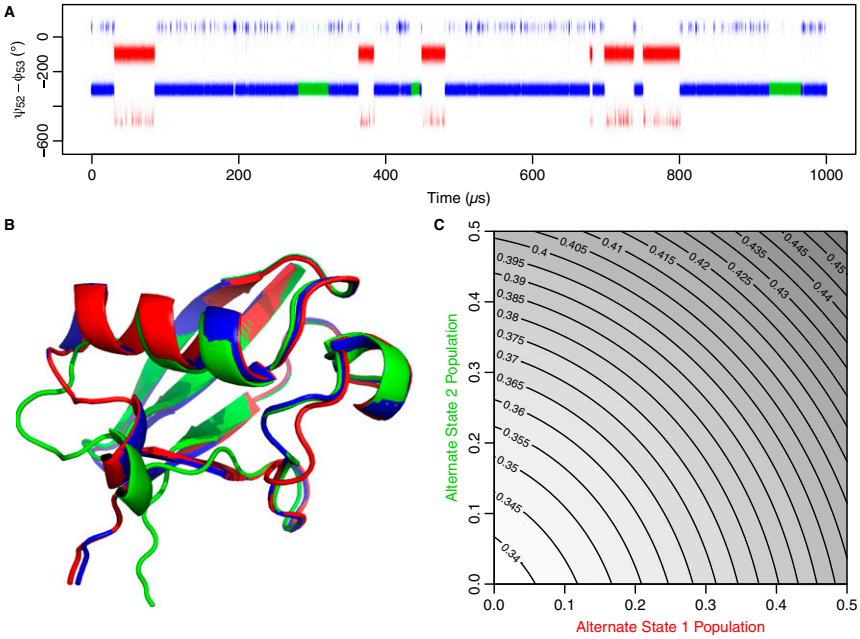


Fig. S6. Peptide flipping is observed in a 1-μs CHARMM22* trajectory. (*A*) Numerous peptide flips between NH-out (−320°) and NH-in (80°) are observed in the dominant state of the trajectory (blue). The peptide flip is also observed in two alternate states (red and blue). (*B*) Dominant state (blue) is similar to all known crystal structures and is 70% populated. The first alternate state corresponds to a local structural change around residues 50–54 (red, 20%). The second involves unfolding of the last turn of the α-helix (blue, 10%). This unfolding may be responsible for the $^{1}H^{N}$ RD observed at I36. (*C*) Population of at least one of these states is thought to be overestimated (22), which is in agreement with our own analysis showing that RDC data are best fit when alternate states 1 and 2 are both assigned a population of zero. Contour lines give the respective Q-factors. Like the AMBER ff99SB trajectories, the rates of peptide bond flipping are about 100-fold faster than in the experiments, likely due to similar inaccuracies in the underlying force field of around 1–3 kcal/mol.

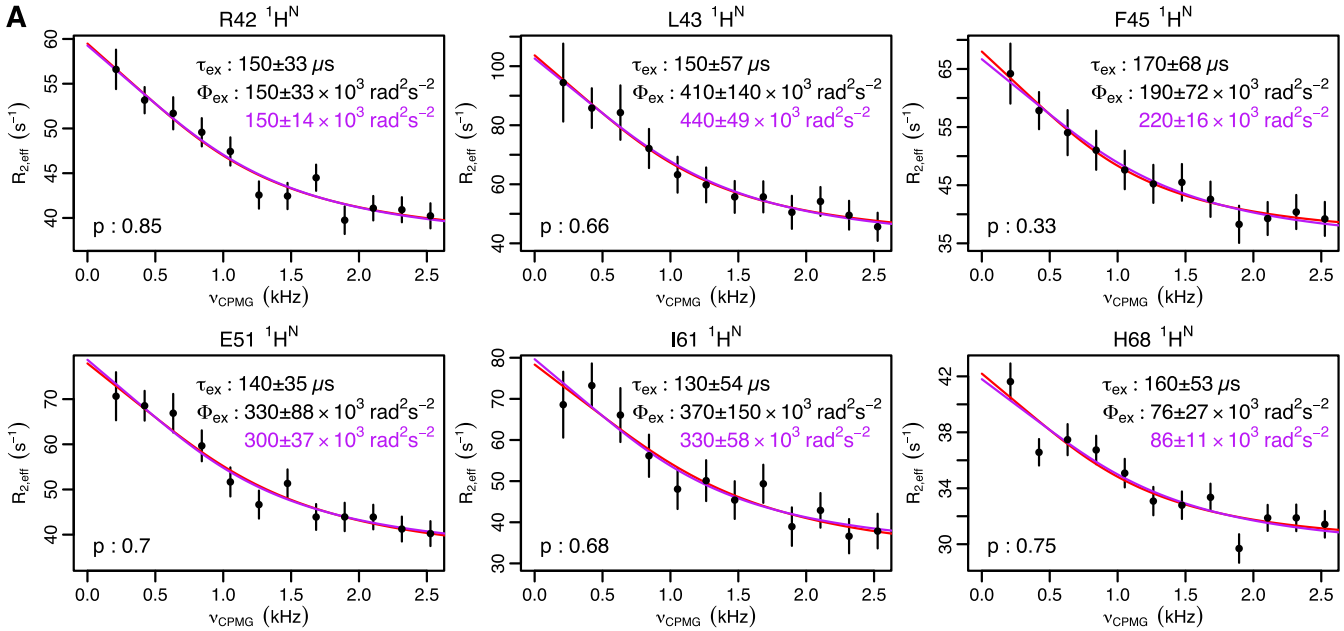


Fig. S7. (Continued)

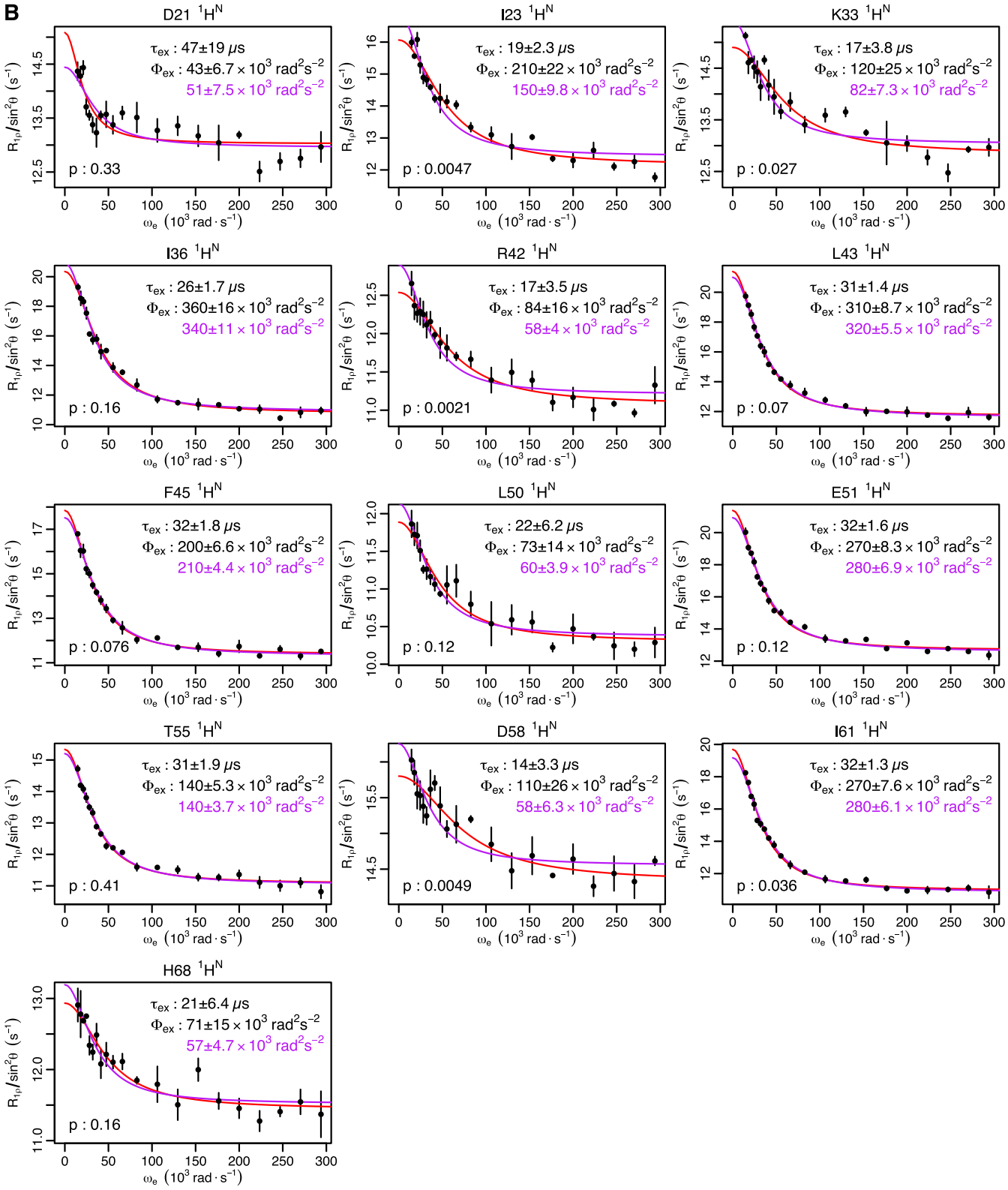


Fig. S7. (Continued)

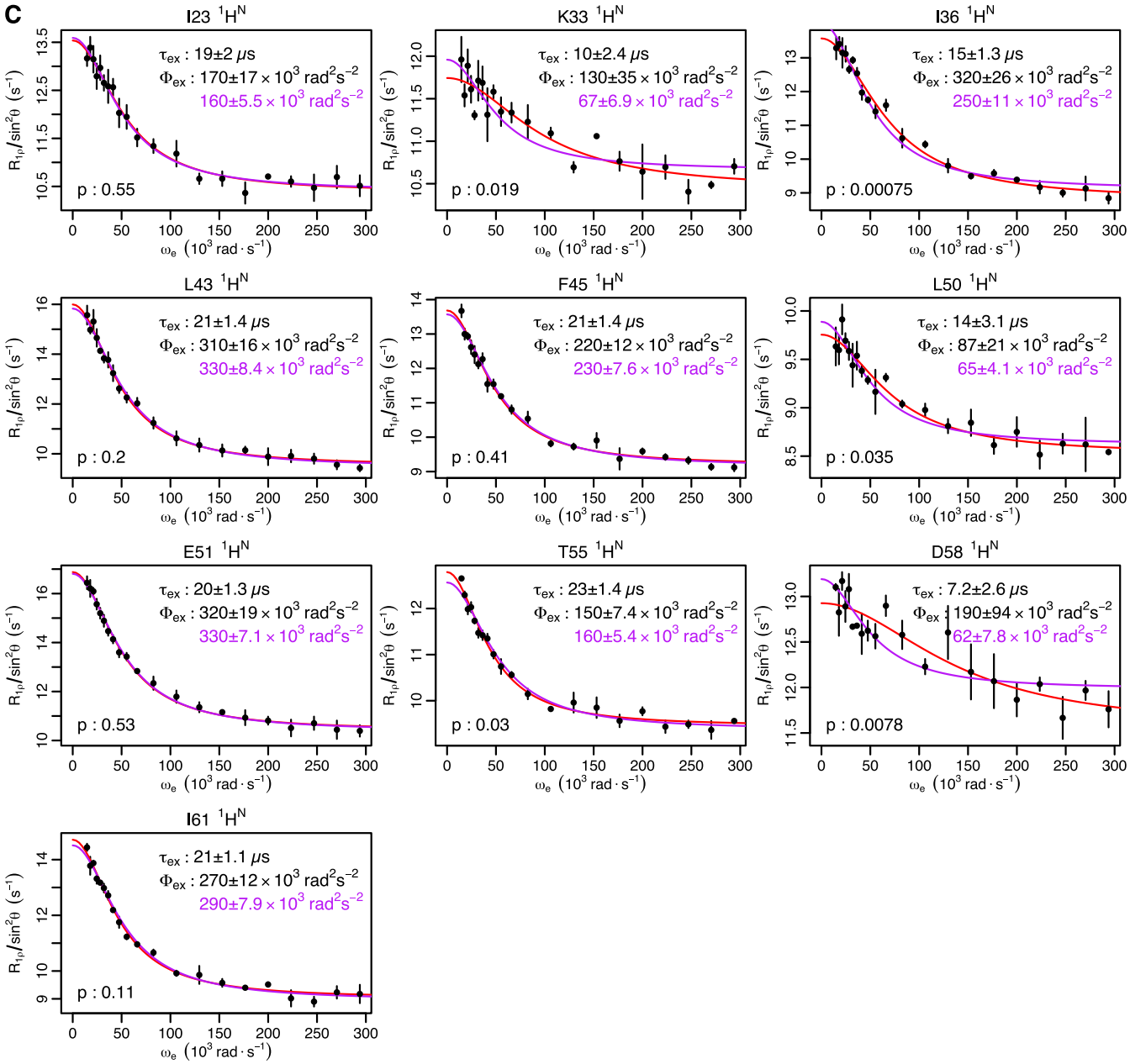


Fig. S7. (Continued)

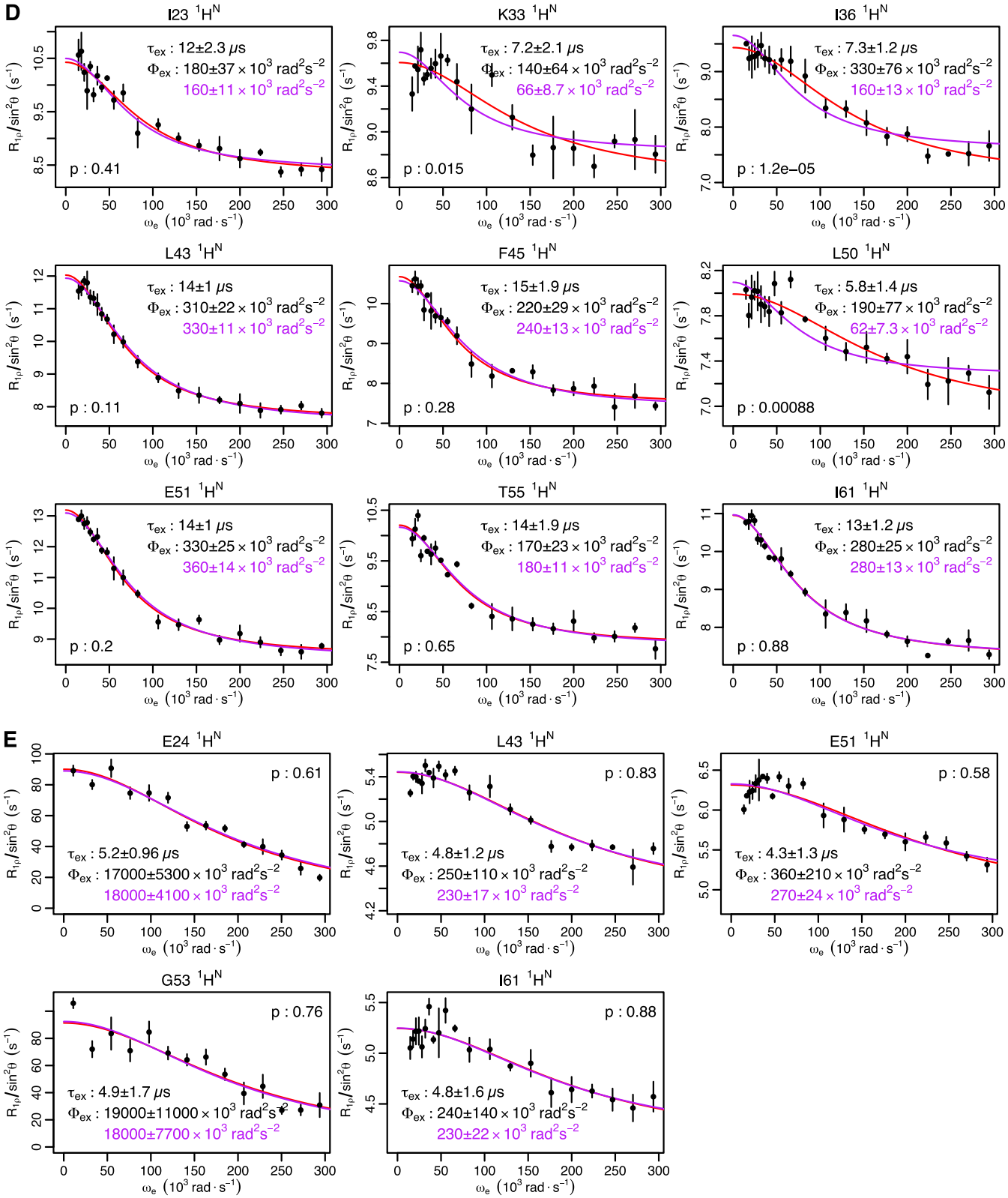


Fig. S7. Ubiquitin CPMG and $R_{1\rho}$ data at other temperatures. Individual fits are shown in red, with the parameters shown in black. Global fits with a single τ_{ex} value are shown, along with the corresponding Φ_{ex} value in purple. *F* test *P* values between fits are also shown. (*A*) Backbone $^1H^N$ CPMG at 262 K (global τ_{ex} = 150 μs). (*B*) Backbone $^1H^N$ $R_{1\rho}$ at 282 K (global τ_{ex} = 29 μs). (*C*) Backbone $^1H^N$ $R_{1\rho}$ at 287 K (global τ_{ex} = 20 μs). (*D*) Backbone $^1H^N$ $R_{1\rho}$ at 292 K (global τ_{ex} = 13 μs). (*E*) Backbone $^1H^N$ $R_{1\rho}$ at 308 K (global τ_{ex} = 5.0 μs).

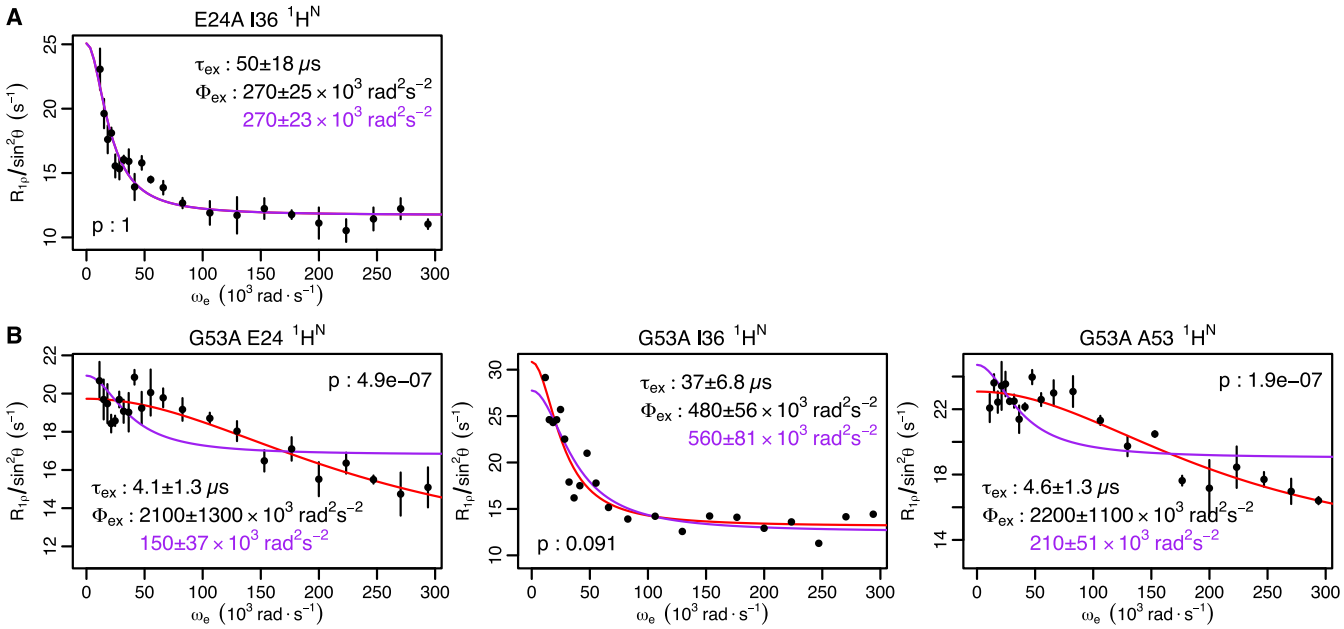


Fig. S8. Ubiquitin mutant $R_{1\rho}$ data. Individual fits are shown in red, with the parameters shown in black. Global fits with a single τ_{ex} value are shown, along with the corresponding Φ_{ex} value in purple. *F* test *P* values between fits are also shown. (*A*) E24A backbone $^1H^N$ of I36 at 277 K (global τ_{ex} = 50 μs). (*B*) G53A backbone $^1H^N$ at 277K (global τ_{ex} = 27 μs). Ten of 11 residues where RD is observed in WT (Fig. S1*B*) do not show RD in the mutants. I36 still shows RD for both mutants, suggesting it reports a different process, possibly unwinding of the α-helix that has been observed experimentally (23, 24) and in simulations (22). The I36 $^1H^N$ time scales are consistent for WT (τ_{ex} = 48 ± 6 μs) and both mutants. For the G53A mutant, the RD at E24 and A53 is about 10-fold faster than WT (55 μs). The Φ_{ex} value for these two nuclei is reduced about 10-fold over the Φ_{ex} value observed at 308 K (Fig. S7*E*), suggesting that the population of the NH-in state is reduced ∼20-fold over WT. A similar 10-fold reduction of the Φ_{ex} values observed for WT at 277 K would drop the 10 residues not observed in the G53A mutant below the detection threshold, explaining their disappearance.

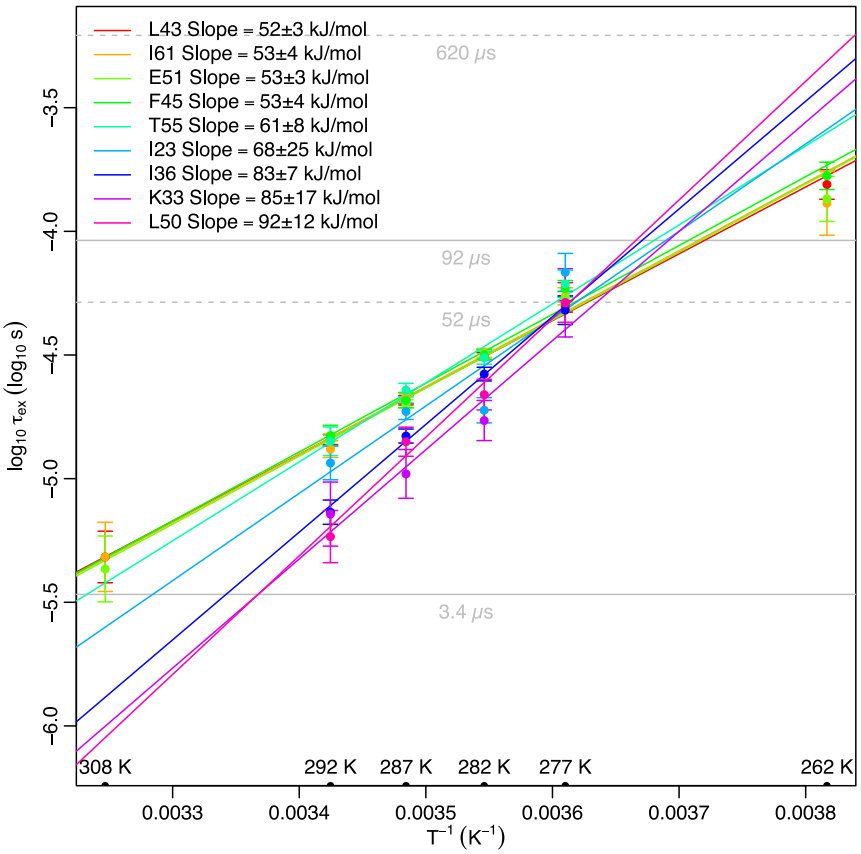


Fig. S9. Temperature dependence of RD time scales. L43, I61, E51, F45, T55, and I23 all show the same temperature dependence, within error. At 308 K, the time scales of L43, E51, and I61 coincide with E24 and G53 (Fig. S7*E*), supporting a direct linkage between all these residues. The temperature dependence of I36, K33, and L50 is somewhat perturbed, suggesting that they may report, in full or in part, on some other process. For I36, this interpretation is supported by mutational data (Fig. 2*E*). Solid and dashed gray lines indicate the accessible time scales for the $R_{1\rho}$ and CPMG experiments, respectively.

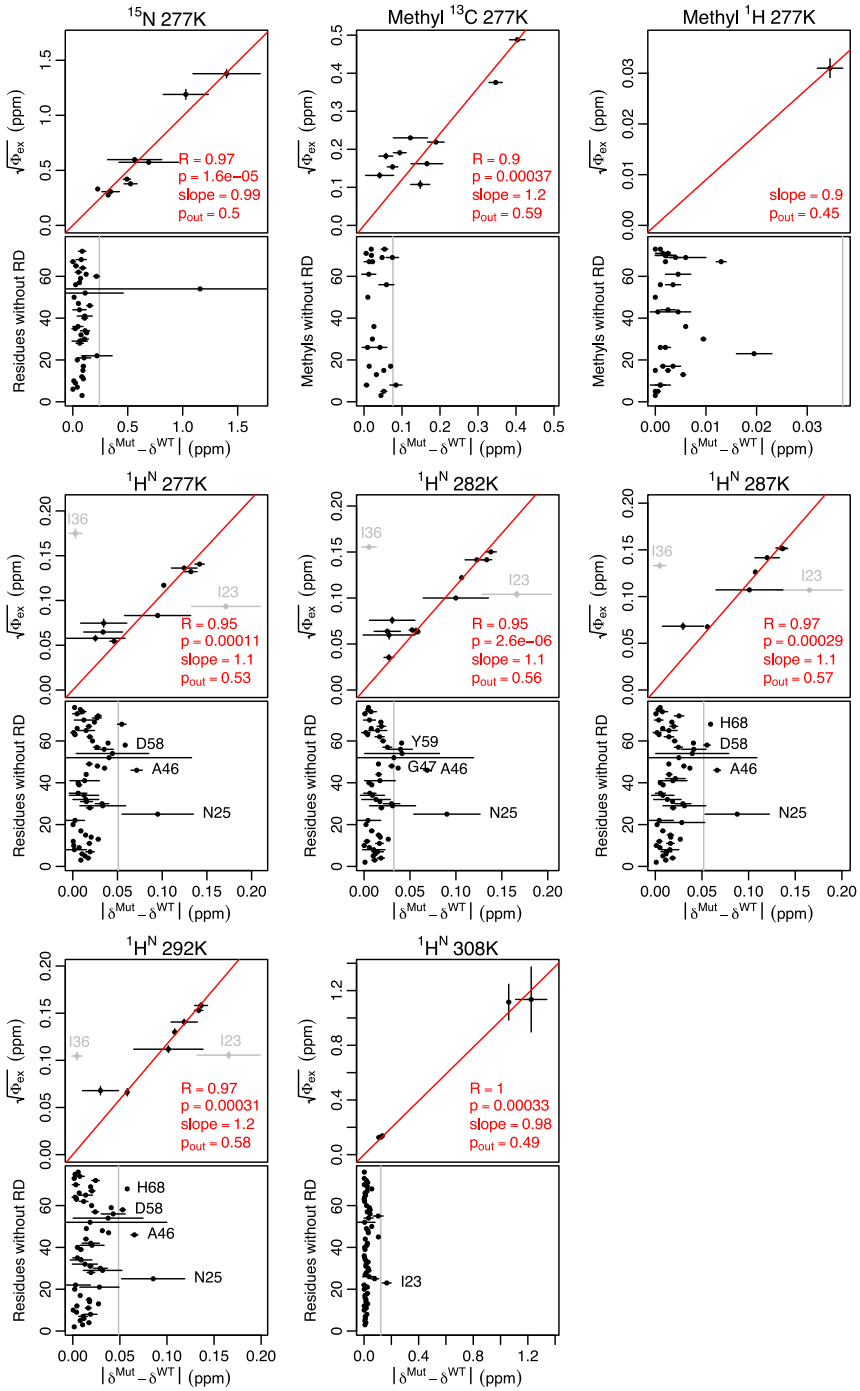


Fig. S10. Mutant chemical shifts explain backbone RD and give state populations. ^{15}N, methyl ^{13}C, and ^{1}H^{N} chemical shift differences between mutant (E24A and G53A) and WT ($|\delta^{Mut} - \delta^{WT}|$) are highly correlated with the observed chemical shift fluctuations from RD ($\sqrt{\Phi_{ex}}$) at all temperatures. The δ^{Mut} values are taken from the mean of the E24A and G53A chemical shifts, with the range of the two mutants covered by the error bars. The Φ_{ex} values are taken from the global fits. Each plot includes Pearson's product moment correlation coefficient (*R*), along with the probability of observing such a value at random if the two parameters were uncorrelated (*p*). Assuming the mutants are entirely in the NH-out state, it can be shown that $p_{out} = 1/((\delta^{Mut} - \delta^{WT})^2/\Phi_{ex} + 1) = 1/(1/slope^2 + 1)$. This equation indicates that a slope of 1 will yield a p_{out} of 0.5. The listed populations come from this equation. Because there is only one methyl ^{1}H data point, only the slope and population are shown. Because ^{1}H^{N} I36 reports on a different exchange process, it was excluded from correlation and population analysis. ^{1}H^{N} I23 was also excluded because it is very close to both mutations and likely includes chemical shift changes due to the mutations themselves and not just the peptide conformation. At each temperature, the lower plot gives the chemical shift changes for residues where it was possible to obtain RD data but significant RD was not observed. Labeled residues have error bars that do not overlap with the error bar of the residue showing the smallest Φ_{ex} (vertical gray line). A46, D58, Y59, and H68 all show RD, but the error margins on the parameters are too large to be considered significant (data not shown). N25 does not show RD, but it is also very close to the mutations and likely to include chemical shift changes directly from the mutations.

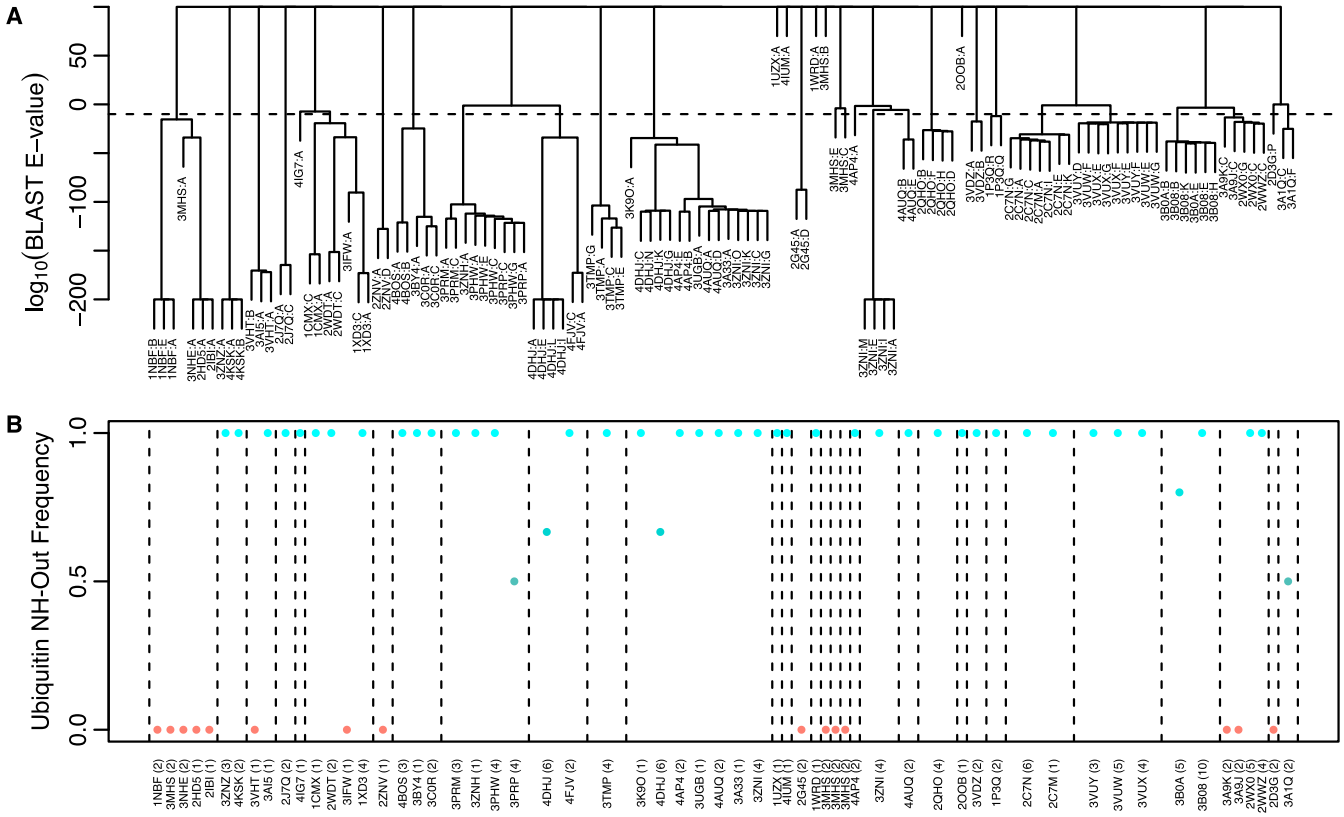


Fig. S11. Complexes with USP deubiquitinases always show NH-in conformations. (*A*) Nonubiquitin sequences from high-resolution (<2.4 Å) structures of ubiquitin complexes were extracted and clustered using the $\log_{10}$ of their BLAST E-values. The PDB identifier and chain containing each nonubiquitin sequence are shown. Groups of sequences were generated using a $\log_{10}$ E-value cutoff of −10. (*B*) For the PDB structures in each group, all unique sets of ubiquitin coordinates were extracted, including all copies in the asymmetrical unit and all alternates. The frequency of finding an NH-out conformation is shown (colored by conformation: entirely NH-in, pink; and entirely NH-out, cyan), with the number of coordinate sets used given in parentheses. The leftmost group is entirely NH-in and corresponds to the USP family of deubiquitinases.

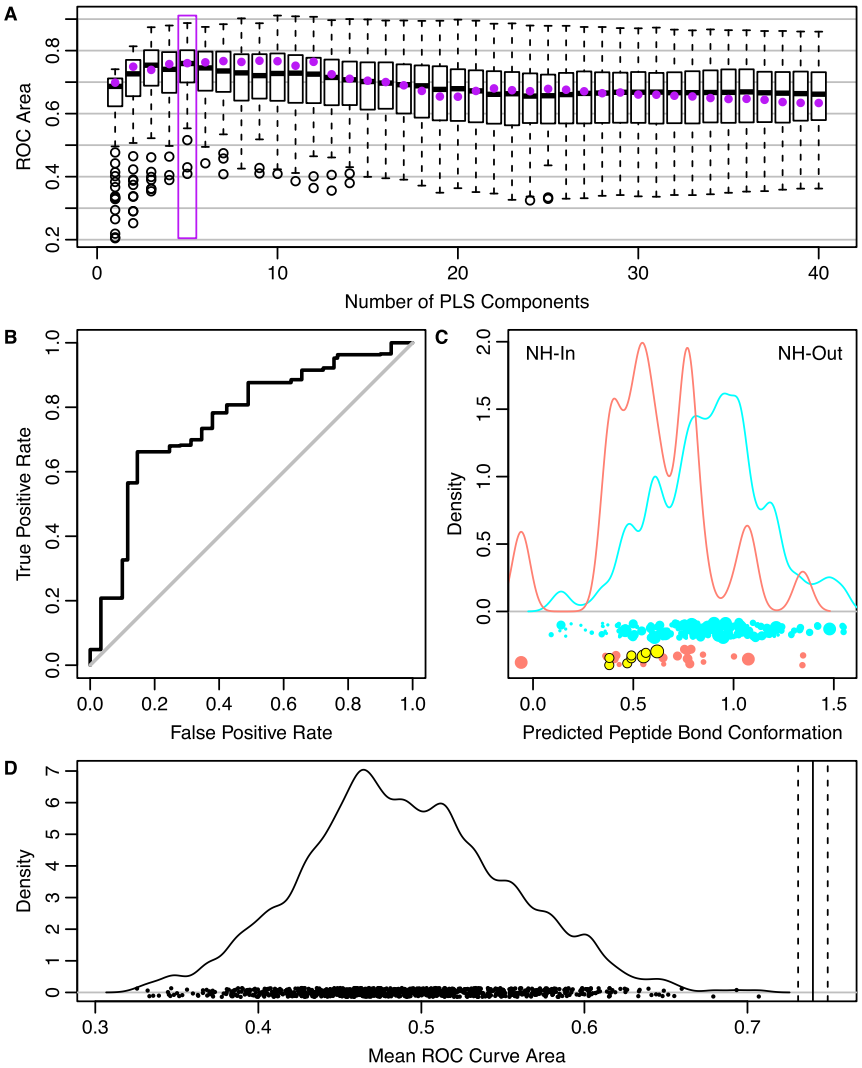


Fig. S12. Cross-validation of peptide fit PDB mode fitting. (*A*) One hundred random twofold cross-validation runs were performed with the constraint that groups shown in Fig. S11*B* not be split between cross-validation groups. For each of these runs, ROC areas were calculated for models that incorporated differing numbers of PLS components. The box plots show the distributions of these ROC areas, with the line giving the median, the box giving the interquartile range, whiskers giving the most extreme data point within 1.5 times the interquartile range, and circles giving data points outside that range. The maximum median ROC area was observed with five PLS components (purple box). This number of components was used in training subsequent models (Fig. 3). A representative cross-validation model was selected whose ROC area was closest to the median ROC area (at five PLS components). All ROC areas for this model are shown with purple points. (*B*) ROC plot for the representative model shows that it is significantly more predictive than random (gray diagonal line). (*C*) Distribution of cross-validated projection values for NH-in (pink) and NH-out (cyan) structures shows moderate overlap between predictions, indicating the USP-interacting residues partially explain the peptide bond conformation. Projection values for USP structures are shown in yellow. In training, the contribution of a set of ubiquitin coordinates was weighted inversely by how many sets of coordinates were in a given PDB structure. The weighting is indicated by the size of the point below the distribution. (*D*) Average cross-validated ROC area (0.74 ± 0.01, solid and dashed vertical lines) is highly unlikely ($P < 0.001$) to have occurred by random chance, based on additional cross-validation runs with shuffled input data. The mean cross-validated ROC curve areas (each from 100 different cross-validation groupings) from 1,000 different sets of shuffled peptide bond conformations are shown as points along the *x* axis. The distribution of these points was determined by kernel density estimation using a Gaussian smoothing kernel with the SD (0.006) taken from the average SE of the 1,000 points. Because none of the shuffled data points exceeds the unshuffled value, the *P* value is estimated as being less than 1/1,000.

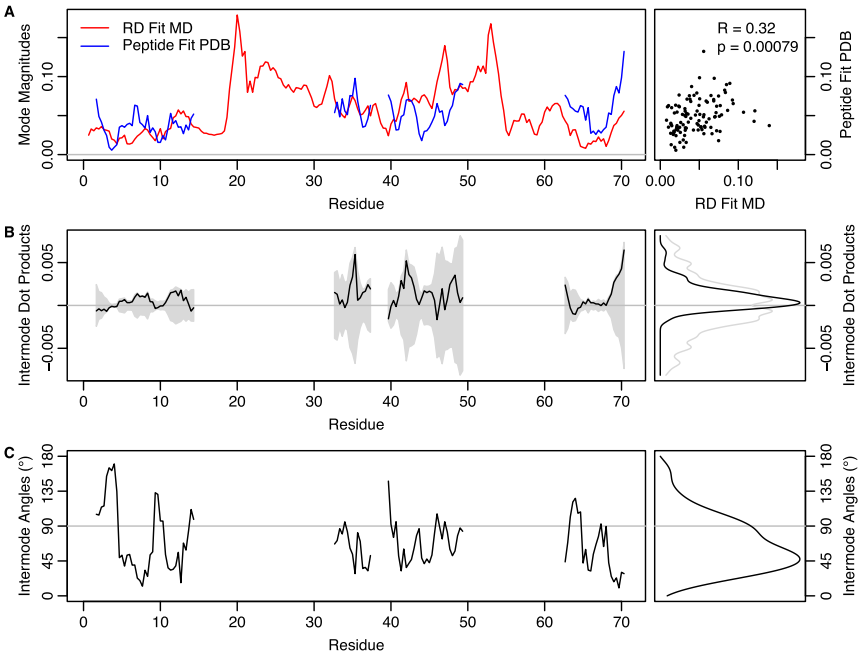
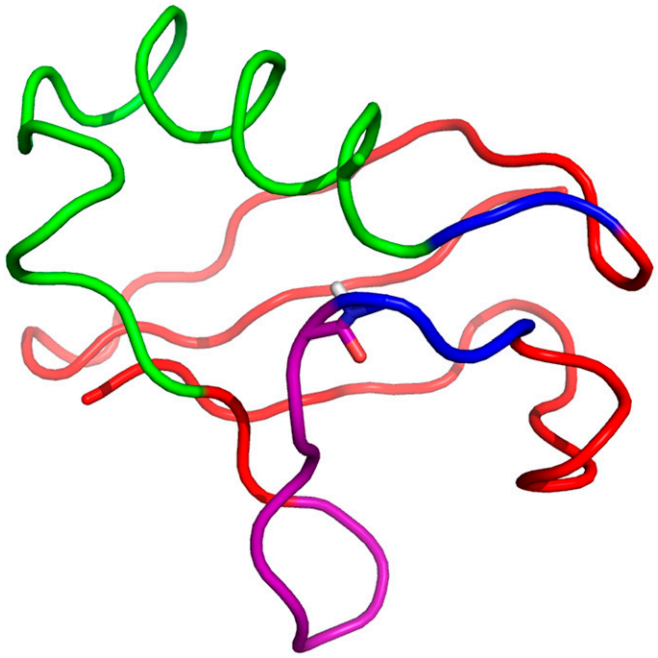
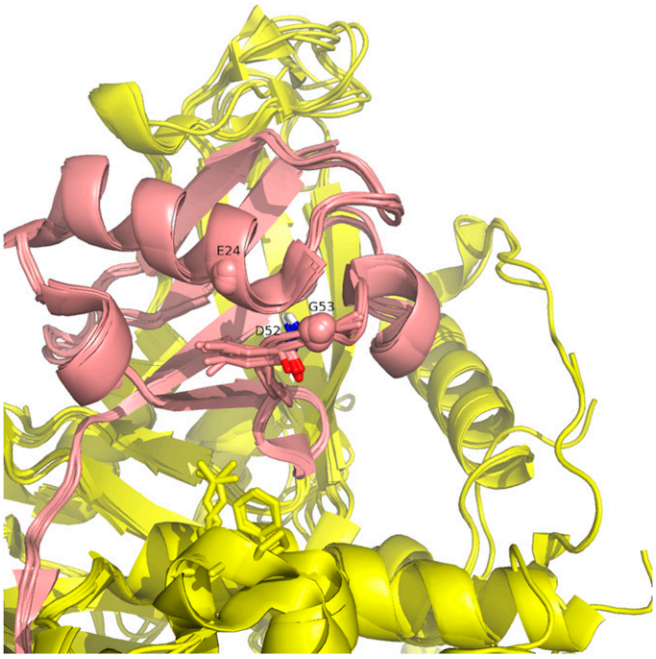


Fig. S13. RD fit MD mode and peptide fit PDB mode are similar. The magnitudes and directions of motion for every atom (backbone N, Cα, C) were extracted from both the RD fit MD mode and peptide fit PDB mode. (*A*) Magnitudes of the RD fit MD mode (red) do coincide with magnitudes of the peptide fit PDB mode (blue). The correlation is modest ($R = 0.32$) but statistically significant ($P = 0.00079$). (*B*) Mostly positive (84 of 108) per-atom dot products between modes indicate that the majority of atoms move in similar directions in both modes (Wilcoxon signed rank: $P = 9.9 \times 10^{-11}$). The shaded area indicates the range of possible per-atom dot products, given the magnitudes shown in *A*. (*C*) Peak angular difference between per-atom directions of motion is ∼45°.



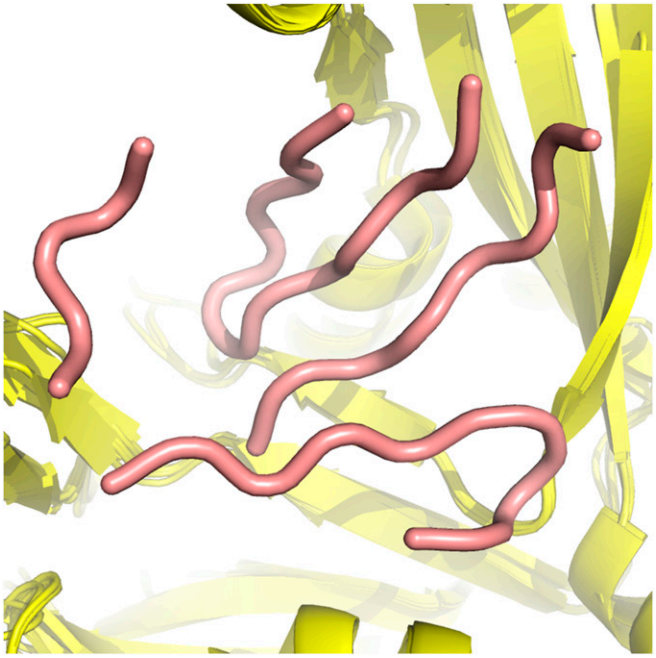
Movie S1. RD fit MD mode. Interpolation of between extremes of the RD fit MD mode is as shown in Fig. 1*F*.

Movie S1



Movie S2. Ubiquitin/USP complex crystal structures. Six high-resolution (<2.4 Å) ubiquitin (pink)/USP (yellow) crystal structures are shown (1NBF:C/B, 1NBF:D/A, 2HD5:B/A, 2IBI:B/A, 3MHS:D/A, and 3NHE:B/A). Ubiquitin residues D52 and G53 are shown with a stick representation. Any USP residue within 10 Å of the D52-G53 peptide bond (colored red, white, and blue) is also shown with a stick representation. The Cα atoms of E24 and G53, which were mutated to Ala, are shown as spheres.

Movie S2



Movie S3. Peptide fit PDB mode. Interpolation of between extremes of the peptide fit PDB mode is as shown in Fig. 4*A*.

Movie S3