Published in final edited form as:
J Chem Theory Comput. 2017 June 13; 13(6): 2418–2429. doi:10.1021/acs.jctc.7b00059.

Determining Atomistic SAXS Models of Tri-Ubiquitin Chains from Bayesian Analysis of Accelerated Molecular Dynamics Simulations

Samuel Bowerman[†,‖], **Ambar S.J.B. Rana**[‡,¶,‖], **Amy Rice**[†], **Grace H. Pham**[¶], **Eric R. Strieter**[*,‡,§], and **Jeff Wereszczynski**[*,†]

[†]Department of Physics and Center for the Molecular Study of Condensed Soft Matter, Illinois Institute of Technology, Chicago, Illinois 60616, United States

[‡]Department of Chemistry, University of Massachusetts–Amherst, Amherst, Massachusetts 01003, United States

[¶]Department of Chemistry, University of Wisconsin–Madison, Madison, Wisconsin 53706, United States

[§]Department of Biochemistry and Molecular Biology, University of Massachusetts–Amherst, Amherst, Massachusetts 01003, United States

Abstract

Small-angle X-ray scattering (SAXS) has become an increasingly popular technique for characterizing the solution ensemble of flexible biomolecules. However, data resulting from SAXS is typically low-dimensional and is therefore difficult to interpret without additional structural knowledge. In theory, molecular dynamics (MD) trajectories can provide this information, but conventional simulations rarely sample the complete ensemble. Here, we demonstrate that accelerated MD simulations can be used to produce higher quality models in shorter time scales than standard simulations, and we present an iterative Bayesian Monte Carlo method that is able to identify multistate ensembles without overfitting. This methodology is applied to several ubiquitin trimers to demonstrate the effect of linkage type on the solution states of the signaling protein. We

[*]**Corresponding Authors:** estrieter@chem.umass.edu. jwereszc@iit.edu.
[‖]**Author Contributions**
S.B. and A.S.J.B.R. contributed equally to this work.

Supporting Information
The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.7b00059.

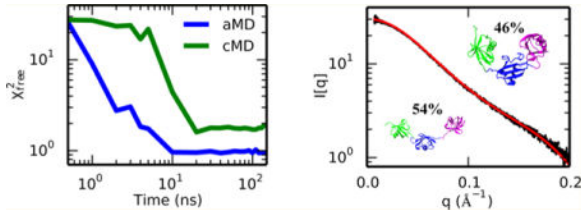
Discussion of AIC using a simplified model, further elaboration of χ^2_{free}, additional details regarding trimer construction, elaboration of aMD details, table of aMD parameters, table containing cMD model results (analogous to Table 2), signal-to-noise ratio vs q-value for each trimer, seven figures of RMSD time series data, number of scattering states vs the number of structural states for the remaining six systems (analogous to Figure 4), model goodness-of-fit convergence for K6-, K11-, and K29-linked trimers, and final models for the K6-, K11-, K29-, K48-, nK48-, and nK63-linked systems (PDF)

ORCID
Eric R. Strieter: 0000-0003-3447-3669
Jeff Wereszczynski: 0000-0002-2218-3827

Notes
The authors declare no competing financial interest.

observe that the linkage site directly affects the solution flexibility of the trimer and theorize that this difference in plasticity contributes to their disparate roles in vivo.

Graphical abstract



1. INTRODUCTION

Characterizing the inherent flexibility of biomolecular complexes remains one of the key challenges in modern biophysics research.[1] Over the years, a wealth of both experimental and computational techniques have been developed that can explore the solution dynamics of proteins on a vast range of time and length scales. One particularly useful method for studying these properties is small-angle X-ray scattering (SAXS).[2,3] Unlike X-ray crystallography, which requires cohesive repeats of a singular representative structure in nonnative conditions, SAXS experiments probe the entire ensemble of a protein in physiologically relevant environments. SAXS experiments are relatively easy to perform and produce data that represent an ensemble average of all states at room temperature with results that are not influenced by factors such as lattice packing forces. Nevertheless, these advantages come at a cost: the primary data resulting from SAXS experiments, the scattering curve, is of low dimensions and can be difficult to interpret.[2] As a result, the analysis of SAXS data typically requires ancillary structural information provided by complementary methods, such as crystallography or spectroscopy,[4–7] which possess their own limitations with regard to system size and intrinsic disorder.[8,9]

In principle, the conformational ensemble of a system can be fully characterized with molecular dynamics (MD) simulations. However, in practice MD is limited by both the amount of sampling that can be performed as well as the accuracy of the underlying models. Recent advances in computing power and enhanced sampling techniques have opened the door to MD-based methods that can probe biomolecular dynamics that occur on the μs-ms time scale.[10–18] For example, in replica exchange methods, multiple copies of the same system are simulated at either different temperatures or with different Hamiltonians, and a Markov Chain Monte Carlo algorithm is used to transfer information between neighboring replicas.[11,12] Metadynamics provides an alternative enhancement method in which the potential energy landscape underlying a system is smoothed through a history-dependent bias, thereby reducing the amount of sampling time required to escape local minima and increasing the conformational space sampled in a simulation.[17,18] Accelerated molecular dynamics (aMD) speeds sampling through the introduction of a "boost" potential that reduces the depths of energy wells and thus increases the rate of barrier crossings.[10,19] Unlike methods such as Hamiltonian replica exchange or metadynamics, aMD does not require the definition of a reaction coordinate along which to bias, and acceleration can be achieved without performing multiple concurrent simulations. On the other hand, aMD does disturb the Boltzmann distribution of states observed in a simulation and calculating the physically relevant populations of states from an aMD simulation can be difficult, especially for large biomolecular complexes.

The relative strengths and weaknesses of SAXS and enhanced sampling methods make them a natural complement to one another.[20] Indeed, several tools have been developed to directly use the high-resolution structures from enhanced sampling simulations to produce an ensemble-based interpretation of low-resolution SAXS data. For example, the SASSIE Web server conducts Monte Carlo samplings of molecular torsions and subsequently evaluates scattering states in a single interface.[21–24] In the BILBOMD protocol, multiple timestepping algorithms are coupled with high temperatures, and observed structures are subsequently reweighted using a genetic algorithm.[25] In another approach, collective modes gathered from an elastic network model of a protein are amplified during MD simulations, and representative structures are fit by the genetic algorithm implemented in the Ensemble Optimization Method of the ATSAS package.[26,27] A fourth method, BSS-SAXS, utilizes coarse-grained simulations to form a scattering basis and then distributes the ensemble population through a Bayesian Monte Carlo reweighting scheme.[28]

These methods are particularly useful for flexible molecules where a single conformation cannot sufficiently describe the observed profiles. However, further conversation is required with regard to overfitting multistate models. That is to say, the parameters of an ensemble model can create arbitrarily strong goodness-of-fit values through the introduction of too many subpopulations. This can be viewed as a violation of Occam's Razor,[29] where the addition of extra scattering states corresponds to an overly complicated model that contains irrelevant components. Therefore, an accurate model ensemble must resist this overfitting tendency and instead report the minimum number of states required to achieve good agreement with experimental scattering data.

Ubiquitin chains represent prototypical examples of highly flexible systems that are difficult to characterize through SAXS or MD methods alone. Within these oligomers, ubiquitin units are linked to one another through an isopeptide bond that forms between the C-terminus of one monomer and the ε-amino group of one of seven ubiquitin lysines (K6, K11, K27, K29, K33, K48, and K63; see Figure 1) or the N-terminus (M1). The dogma is that chain linkages dictate biological function. For example, K48-linked chains are typically involved in proteasomal degradation,[30] whereas K63-linked chains act as scaffolds for the formation of multiprotein complexes.[31,32] In vitro, ubiquitin subunits may be connected to one another through non-native linkages, such as with thiol–ene coupling, to quickly and robustly generate libraries of diverse chain linkages.[33,34] These non-native linkages generally have similar structural properties as those of native isopeptide bonds; however, they can be significantly easier to produce.[35]

Here, we used SAXS experiments along with enhanced sampling MD simulations to characterize the conformational ensemble of seven ubiquitin chains. To do this, we developed a modified BSS-SAXS approach in which cMD and aMD simulations are combined with an iterative Bayesian population reweighting approach to rapidly model

SAXS data. Our methods place an emphasis on determining ensembles that simultaneously minimize goodness-of-fit while resisting overfitting of the data. Results are compared for both conventional MD (cMD) and aMD simulations, and it was found that cMD simulations never outperformed the aMD trajectories in model quality or convergence speed. In fact, some aMD simulations produced better models up to an order of magnitude faster than cMD simulations of the same system. In all, these systems show that aMD enhanced sampling provides a straightforward method for quickly producing robust all-atom models from SAXS experiments.

2. METHODS

2.1. Experimental Details

2.1.1. Expression and Purification of Ubiquitin Variants—Lysine to cysteine mutations were introduced at specified sites in the DNA sequence of ubiquitin (Ub 1–76) using splice overlap extension. Primers containing the TGC mutation were inserted at the desired codon position. A codon extending the C-terminus of ubiquitin with D77 was introduced using reverse primer to afford all five constructs (UbK*x*C-D77, where *x* represents the position of the native lysine residue that was replaced), and then they were ligated into a pET22b vector (Novagen). Wild-type ubiquitin and its variants were expressed and purified from Rosetta 2(DE3)pLysS cells (Novagen) as previously described.[33,35]

2.1.2. Generation of Native and Non-Native Ubiquitin Trimers—As previously described, the yeast C-terminal hydrolase Yuh1 was used to append allylamine (AA) to the C-terminus of ubiquitin to afford UbKxC-AA variants.[33] Irradiation of UbKxC-AA (2 mM) lithium acyl phosphinate (LAP) (0.5 mM) in 250 mM NaOAc pH 5 for 30 min at 4 °C yielded UbKxC oligomers with non-native thiol–ene-derived linkages.[33,35] All ubiquitin trimers were purified using size exclusion chromatography (Hiload 26/600 Superdex 75 pg, GE Healthcare).

2.1.3. SEC-SAXS Measurements and Data Processing—Size-exclusion chromatography small-angle X-ray scattering (SEC-SAXS)[37] experiments were performed at BioCAT (beamline 18-ID, Advanced Photon Source at Argonne National Laboratories). The camera included a focused 12 keV (1.03 Å) X-ray beam, a 1.5 mm quartz capillary sample cell, and a Pilatus 1 M detector (Switzerland). The q-range sampled was ~0.0045–0.35 $Å^{-1}$. To ensure sample monodispersity, we used an in-line SEC setup, which included an AKTA-pure FPLC unit and a Superdex-75 10/300 GL column (GE Healthcare Life Sciences). The column was run at 0.8 mL/min, and the outlet was directly connected to the SAXS sample cell. One second exposures were collected every two seconds during the gel-filtration chromatography run. Samples were analyzed at room temperature in 50 mM HEPES pH 7.5, 50 mM KCl, 5 mM $MgCl_2$, and 1 mM TCEP. Exposures before and after the elution of the sample were averaged and used as the buffer blank curve, and the exposures during elution (coincident with the UV peak on the chromatogram) were treated as protein plus buffer curves. Data were corrected for background scattering by subtracting the buffer blank curve from protein plus buffer curves. The radius of gyration (R_g) for each system was determined with the aid of PRIMUS,[38] and the resulting R_g values obeyed the limitation

$q_{\mathrm{max}} \cdot R_{\mathrm{g}} < 1.3$. To minimize the impact of low-q beamstop effects on model quality, experimental data were filtered such that all low-q points with a signal-to-noise ratio of 10.0 or below were manually removed and not used in the fitting protocol. Data were also truncated at q = 0.2 Å^{-1} due to the inherent limitation of implicit hydration layer SAXS calculations.[39]

2.2. Simulation Details

2.2.1. Construction of the Triubiquitin Systems—Initial structures were obtained from the Protein Data Bank for monoubiquitin (PDB: 1UBQ[36]), K11-linked diubiquitin (PDB: 3NOB[40]), K6-linked diubiquitin (PDB: 2XK5[41]), K29-linked diubiquitin (PDB: 4S22[42]), K48-linked diubiquitin extracted from K48-linked cyclic tetraubiquitin (PDB: 3ALB[43]), and K63-linked tetraubiquitin (PDB: 3HM3[44]). Any missing or mutated residues in these structures were corrected by comparison with the 1UBQ structure of monoubiquitin with missing coordinates taken by aligning 1UBQ with the monomer of interest. K63-linked triubiquitin was constructed by removing one monomer from the tetraubiquitin crystal structure. The K6-, K29-, and K48-linked systems were constructed by overlapping two copies of the corresponding diubiquitin structures and then removing the overlapped monomer. Finally, the K11-linked triubiquitin was constructed by adding the monoubiquitin structure to the K11-linked diubiquitin crystal structure. Table S1 provides an overview of the PDB structures used to construct each system.

To match the prepared ubiquitin samples, all five linkage types were simulated with a non-native thiolene linkage between monomers instead of the native isopeptide bond (Figure 2). Additionally, two of the systems, K48 and K63, were also simulated with native isopeptide linkage to determine how linkage chemistry affects molecule dynamics and the solution ensemble. Both sets of linkage parameters were generated with GAFF with partial charges derived from a restrained electrostatic potential (RESP) fit to the electrostatic potential computed at the HF/6-31G* level using Gaussian 09.[45] Each system was solvated in a box of TIP3P water[46] with 12 Å of padding in each dimension; then, sodium and chloride ions were added to a 0.15 M concentration. The Amber force field ff14SB was used[47] along with the isopeptide and thiolene modifications described above.

2.2.2. Conventional Molecular Dynamics—Initial equilibration was performed at constant temperature in both the constant volume and constant pressure ensembles with a 2 fs time step. Before the equilibration phase, the solvent was energy minimized using both the steepest descent and conjugate gradient algorithms. In the equilibration phase, the system was first heated in the NVT ensemble to 310 K over 50 ps with restraints on all non-hydrogen atoms of the solute. Next, the systems were simulated for 500 ps in the NPT ensemble with restraints on all non-hydrogen atoms of the solute to allow the density of the solvent to equilibrate. Finally, the restraints were slowly relaxed over 3 ns. No restraints were applied during the production phase, which was carried out for 500 ns in the NPT ensemble. Frames were saved every 5 ps during the production phase.

The temperature of the system was maintained at 310 K using Langevin dynamics with a collision frequency of 1.0 ps^{-1}. The pressure was maintained at 1.0 bar by means of isotropic coordinate scaling utilizing the Berendsen barostat and a relaxation time of 1.0 ps.[48] All hydrogen bonds were constrained using SHAKE.[49] A cutoff of 10.0 Å was used for direct nonbonded interactions, and long-range electrostatics were treated with the particle mesh Ewald (PME) method with a 1 Å grid spacing.[50] All simulations were performed with the gpu-accelerated version of *pmemd* in Amber 14,[13,51] and R_g time series were calculated using cpptraj.[52]

2.2.3. Accelerated Molecular Dynamics—Following the minimization and equilibration phase simulations, each system was also simulated in the NPT ensemble using accelerated molecular dynamics (aMD).[10] The aMD simulations were performed using the dual-boost variant[19] in which a boost potential is applied to the whole potential with an extra boost to the torsions. The boost is defined by four aMD parameters, E_d, α_d, E_p, and α_p, where $E_{p/d}$ defines the energy threshold below which the boost should be supplied and $\alpha_{p/d}$ is the acceleration factor that controls the shape of the modified potential. Values for E_d, α_d, E_p, and α_p were calculated individually for each system based on the number of atoms, dihedrals, and average energies from the first 50 ns of the cMD simulations (see Table S1).[53] For K6-linked triubiquitin, these parameters were insufficient to attain the desired sampling level (see section S4 for further discussion), so E_p and E_d were each incremented by the value of the corresponding α. All aMD simulations were carried out for 150 ns using the same simulation parameters as described above. Lastly, aMD ensembles were not reweighted according to the Boltzmann weight of the applied boosts, as populations of states were extracted directly from fitting to SAXS data after the simulation was completed.

2.3. Bayesian Modeling of SAXS Data

We developed and utilized an iterative Bayesian reweighting scheme to determine the ensemble of structures from our simulations that best represented the solution SAXS measurements (Figure 3). In summary, MD trajectories were subjected to two rounds of clustering: one based on their structures and another based on their theoretical scatting profiles. Structural clustering allowed us to focus on atomic scale structures that were sufficiently unique from one another, and the scattering clustering step allowed us to only consider structures that were experimentally distinguishable from one another. Ensembles of increasing size of theoretical scattering curves were then considered with a Bayesian Monte Carlo algorithm until overfitting was observed, at which point the statistically best-fit model was chosen.

2.3.1. Creation of Scattering Clusters—For decorrelating the configurations sampled in our MD simulations, N structural states were isolated using an agglomerative clustering algorithm.[54,55] For each structural state, a theoretical SAXS scattering profile was calculated by multipole expansion.[56] For the K nonredundant and experimentally distinguishable SAXS profiles from these N scattering curves to be determined, another agglomerative clustering was then utilized with the distance between scattering profiles metric defined as

$$S_{i,j}=\frac{1}{n_s}\sum_{q\varepsilon N_s}\left(\frac{I_i(q)-I_j(q)}{\sigma(q)}\right)^2 \qquad (1)$$

where $I_i(q)$ and $I_j(q)$ are the scattering intensities of profiles i and j at momentum transfer q, $\sigma(q)$ is the experimental scattering error, and N_s is the number of data points. The summation was subsampled based on the Nyquist–Shannon sampling theorem to ensure that data points in the scattering profiles were decorrelated from one another.[57] In this metric, an $S_{i,j}$ of 1 suggests that two profiles are indistinguishable within the experimentally observed noise. In this way, the 2σ approach to signal recognition corresponds to clustering the similarity scores to a value of $S_{i,j} = 2$.

2.3.2. Bayesian-Based Population Estimates—The theoretical scattering profile for the ensemble of K scattering states is calculated as the population weighted average of the individual states[27,28]

$$I_t(q)=\langle I(q)\rangle=\sum_{i=1}^{K} w_i I_i(q) \qquad (2)$$

where $I_i(q)$ is the scattering profile of state i and w_i is its fraction of the total population. Although multiple well-established techniques have been developed to calculate $I_i(q)$ from atomic coordinates,[56,58–60] determining the population of states w_i is a nontrivial task. Here, we employed a modified and iterative implementation of the BSS-SAXS[28] approach to determine the population of states.

The posterior distribution of Bayes’ theorem is defined as[61]

$$p(\theta|X)=\frac{p(X|\theta)p(\theta)}{p(X)} \qquad (3)$$

where X represents the experimental scattering profile ($I(q)$), θ is the set of population weights ($\{w_i\}$), $p(X)$ is the marginal likelihood of the data, $p(\theta)$ is the prior distribution (probability distribution for the set of population weights $\{w_i\}$), $p(X|\theta)$ is the likelihood function (the probability that the set of weights $\{w_i\}$ can reproduce the experimental scattering profile), and $p(\theta|X)$ is the posterior distribution from which the set of weights $\{w_i\}$ that best models the experimental data are extracted. The likelihood function $\exp(-\chi^2_{\text{free}}/2)$ is employed because experimental SAXS errors are approximately Gaussian,[62–64] but model qualities are reported as reduced χ^2_{free} values, as this is the colloquial value used when reporting SAXS models. The χ^2_{free} metric is used in place of the standard χ^2 as it is a more accurate measure of model quality and less prone to overfitting (section S2).[63] A reduced χ^2_{free} can be calculated analogously to the reduced χ^2 metric, but the nonreduced form was used in the likelihood function. In theory, a prior distribution with many features could be calculated from the MD trajectory. However, the sampling of the simulations presented here was sufficient to explore a variety of conformations but insufficient to discern their relative populations of states. Therefore, the prior distribution of population weights was assumed to be uniform.

A Baysian Monte Carlo algorithm was used to explore the θ parameter space and subsequently map the posterior distribution. For overfitting through an excessive number of possible scattering states to be avoided, an iterative "bottom-up" method was employed to determine the minimal number of structures required. First, the goodness-of-fit for each individual scattering state was determined, along with an associated Akaike information criterion (AIC)[65]

$$\mathrm{AIC}{=}2\nu - 2\ln(\hat{L}) \quad (4)$$

where ν is the number of model parameters and $\hat{L}$ is the maximum observed likelihood (i.e., the minimum observed χ^2_{free}). The AIC penalizes a model's goodness-of-fit from arbitrarily increasing the number of model parameters (see section S1 of the Supporting Information for more details). Following the inspection of all single states, all permutations of ensembles that contained two states were considered. This procedure was repeated with incrementally increasing basis set sizes until the model AIC was not improved. Following these permutations, the minimum AIC model was reported as the one that best fits the experimental data.

2.3.3. Implementation—Initially, the MD ensembles from each simulation were structurally grouped using the hierarchical clustering protocol implemented in cpptraj.[52] Coordinates from every 20 ps of simulation were least-squares fit according to their backbone atoms, and distances between members were defined using the RMSD of Cα atoms. To determine the effect of structural clustering on the final number of scattering states, several different structural bases were produced by defining the total number of clusters (N) to be 25, 50, 100, 200, 300, and 500. For each structural basis, the central member of each cluster was selected as the representative state, and the scattering profile of each representative was calculated using Crysol.[56] Structural clusters were then further grouped into scattering states separated by a similarity score of $S_{i,j}$ = 2. Most systems did not display a large deviation in scattering numbers when more than 300 initial structures were considered, so N = 300 initial structural clusters was chosen for all systems (see below). For the sampling convergence calculations, if the total number of frames was less than or equal to 300, then no structural clustering was conducted, and the K scattering states were clustered from the theoretical profiles of every frame.

The number of Shannon channels for determining χ^2_{free} was defined using the SHANUM program of the ATSAS suite.[66] For each basis subset permutation, ten randomly initiated Monte Carlo searches were conducted for a total of 10,000 steps, and the last 9,000 steps of each run were combined and normalized to create the observed posterior distribution. The population of state i was defined as the average of the marginal posterior in w_i, and the uncertainty was defined as the standard deviation.

3. RESULTS

3.1. SAXS Experiments

Experimental data were gathered for seven different triubiquitin systems, five of which contained non-native thiolene linkages (K6, K11, K29, K48, and K63) and two of which contained native isopeptide bonds (nK48 and nK63). Beam smearing effects were observed at low q, but the SAXS intensity was otherwise well-resolved for each system (Figure S2). Differences in both the scattering curves and R_g values suggest these systems adopt a range of shapes and sizes, and differences in the values of each system suggest varying shapes and sizes based on linkage type (Table 1) with K11 linkages being the most compact and K63 the most extended.

3.2. aMD and cMD Sampling Inefficiencies

The inefficacy of the aMD and cMD simulations to accurately sample the experimental ensemble without reweighting the population of states can be established by comparing the R_g values of each trajectory (columns two and three of Table 1) with experimental measurements. The cMD-based R_g calculations are too compact in five systems (K6, K11, nK48, K63, and nK63), and only the R_g value of the K11-linked trimer is within 1.0 Å of the experimental value. In contrast, the aMD-based predictions are undervalued for every system. Indeed, both simulations are close to being within the error of predicting the R_g of one system: the nK48-linked trimer. However, the R_g predictions of our ensemble-based approach (discussed in detail below) are able to capture nearly every R_g value within the error.

3.3. Identifying Scattering Clusters

For the appropriate method for forming scattering clusters to be determined, the full 150 ns aMD trajectories were analyzed. Each simulation was structurally grouped into a range of total numbers of clusters (N), and then each set of structural representatives was subsequently clustered according to similarity values ($S_{i,j}$) of 1, 2, and 3 (Figure 4 and Figure S10). At each number of structural states, the most scattering states were identified when a similarity score of 1 was enforced, and the fewest number of scattering states were identified for a similarity score of 3. A similarity restriction of 2 was selected because it identified an intermediate number of scattering states and is analogous to a 2σ result in identifying one profile from another. In most simulations, the number of identified scattering states increased with the number of initial structural clusters until $N = 300$, after which there was little change. To test if a larger basis set N would affect our results, the non-native K63-linked system was also clustered according to the scattering profiles of all $N = 7{,}500$ frames and showed the same number of scattering states ($K = 14$) for 7,500 frames as for $N = 300$ structural states.

3.4. Preventing Overfitting with Iterative Inclusion

One inherent difficulty in producing multimember models is the potential of overfitting to experimental data. To demonstrate that our iterative Bayesian approach avoids overfitting, the full basis fitting of the non-native K63 aMD simulation is presented in detail. Initially, the full ensemble of all 14 scattering states was reweighted using a Bayesian Monte Carlo approach (Figure 5). This produced a model with a reduced χ^2_{free} value of 1.12, significantly better than any single scattering state. Although all 14 members were considered in the Monte Carlo, only clusters 2 and 3 contributed more than 0.1 of the population individually and combined for a net 39% of the total ensemble. As a result, the remaining 61% of the population was spread in small amounts throughout the other 12 clusters. This could be the result of overfitting through a large number of model parameters, or it could be the effect of the Monte Carlo sampling along a 14-dimensional hypersurface pooling small amounts of the population to unfavorable states.

For the minimal basis size that best fits the experimental data to be determined, all combinations of basis subsets were considered, and the best-fit model qualities from each subset size were compared (Figure 6). The best single-state model had a modest fit to the experimental data with χ^2_{free} of 2.9. The inclusion of a second conformation drastically improved the fit of our model, lowering the best observed χ^2_{free} to 0.96 for the combination of clusters 2 and 9. This improved goodness-of-fit was met with a decrease in the AIC value from 14.9 in the single-state to 6.5 in the two-state model, justifying the increase in model parameters. In contrast, the optimum three-state model modestly improved the goodness-of-fit $(\chi^2_{\text{free}}{=}0.88)$ but also increased the AIC value to 8.4, suggesting that the benefit of a three-state model in place of the two-state is outweighed by the increase in the number of model parameters. All subsequent basis sizes displayed a consistent increase in AIC value, which suggests that the extra parameters have no beneficial effect on χ^2_{free} reduction and indicates a substantial degree of overfitting.

Surprisingly, clusters 2 and 3, the two most populated members in the full basis ensemble, are not the same combination that form the best two-state model. In fact, a two-state fit using only clusters 2 and 3 resulted in a distribution of nearly the entire population into cluster 3 (reduced $\chi^2_{\text{free}}{=}16.7$), which is significantly worse than the best individual scattering state (reduced $\chi^2_{\text{free}}{=}2.9$), the best two-state model (reduced $\chi^2_{\text{free}}{=}1.0$), and the full 14-member model (reduced $\chi^2_{\text{free}}{=}1.12$). Therefore, the 14-state model appears to represent a drastic overfitting of a poor choice in which two states are most important to the net scattering profile.

3.5. aMD Enhances Model Convergence Relative to that of cMD

Comparisons of aMD and cMD simulations demonstrate that models generated from the accelerated simulations generally converge to lower χ^2_{free} values quicker than their corresponding conventional simulations (Figure 7 and Figure S11). In two cases, K11 and K29, the convergence speed and quality of the aMD models are comparable to those produced by cMD. However, these systems also had the best initial fits $(\chi^2_{\text{free}} \leq 2.5)$, suggesting the comparable performance may be due to the overall quality of the initial conformations.

In contrast, the initial K6 model was quite poor $(\chi^2_{\mathrm{free}} \sim 8)$, and the cMD simulation never converged to an acceptable goodness-of-fit even after 500 ns of simulation (data not shown). However, the aMD simulation converged to an acceptable model $(\chi^2_{\mathrm{free}} \approx 2.5)$ within roughly 70 ns. Combined with the larger deviations in backbone RMSD values in this regime (Figure S3), this system highlights the ability of aMD to sample a wider variety of conformations. Similarly, the nonnative K48 linkage had a large discrepancy between aMD and cMD models (reduced χ^2_{free} of 0.7 and 2.8, respectively). The cMD trajectory identified a single state that is more extended (R_g = 24.3 Å) than is suggested by the experimental data (R_g = 22.3 Å). In contrast, the aMD model accurately predicts the molecular size (R_g = 22.4 Å). Given the trend of the convergence time series (Figure 7), the cMD simulation would likely reach an acceptable solution if the simulation times were significantly extended.

Furthermore, SAXS data were gathered for both native and non-native linkages of K48- and K63-linked trimers. In all four systems, the aMD simulations experienced the best initial improvements to the goodness-of-fit, but the native cMD simulations of both linkage sites eventually converged to the best model in the same time scale as their aMD counterparts. On the other hand, the non-native aMD models outperformed the cMD simulations at all time scales. This disparity between native and non-native convergence is likely due to performing a single simulation of each system and not due to any inherent difference between native and non-native linkages.

3.6. Triubiquitin Ensembles Depend on Linkage Type

Models for all seven triubiquitin systems were produced using both aMD (Table 2 and Figures S12–S17) and cMD (Table S2) simulations. Acceptable models $(\chi^2_{\mathrm{free}} \approx 1.0)$ were found for most systems with the K6-linked cMD simulation having the largest deviation from experimental results at a reduced χ^2_{free} value of 8.7. Additionally, for all but two systems, model ensemble R_g values were within 1 Å of the experimentally determined R_g values (Table 1). Four of the seven systems were best represented by two-state ensembles, whereas the other three were identified as one-state models.

In the two-state models (K11, nK48, K63, and nK63), the final ensemble is generally composed of an open and compact state with the geometry of each varying based on the linkage. The difference between open and closed conformations is significant in the K63, nK63, and nK48 systems with the closed states possessing R_g of ~25 Å and the open states possessing R_g values between 27 and 32 Å. This range in R_g is a direct result of the separation between distal members with the distal groups separated up to 69 Å in the extended states and compacted as low as 27 Å in the closed state. In the K11 system, the difference between open and compact is more subtle with the open and compact states possessing R_g values of 22.5 and 20.7 Å, respectively. Similarly, the distal groups are separated by 35.9 Å in the open and 27.4 Å in the compact states.

The one-state models show a wide variety of geometries based on linkage type. In K6, the best fitting model possessed an R_g of 21.8 Å and a separation distance of 33.4 Å between distal units. In contrast, the molecular size of the K29-linked system was noticeably larger at an R_g of ~25 Å and distal separation of 44.6 Å.

The non-native K48 linkage was also identified as a single state with an R_g of 22.4 Å and a distal separation of 37.1 Å. As previously mentioned, the native K48 linkage was identified as a two-state model, and both the native and non-native K63 linkages were two-state models as well. A comparison of the experimental SAXS profiles from native and non-native K63 shows nearly identical profiles, whereas the non-native K48 is significantly different from its native counterpart, leading to the differences in identified computational models. Together, these results suggest that the solution ensemble of ubiquitin oligomers may be more affected by non-native linkages at some sites than at others.

4. DISCUSSION

Here, we have developed and applied an ensemble fitting method that utilizes MD simulations to fit experimental SAXS data. One critique of ensemble fitting is the potential of overfitting, where a large number of possible states leads to a better goodness-of-fit primarily by increasing the number of model parameters. However, our iterative Bayesian approach avoids this by not assuming the necessity of a particular population size but instead considering the full permutation of increasing subset sizes. By evaluating the AIC value of each ensemble, models that benefit solely from an increasing parameter space are correctly rejected in favor of smaller population sets, as is shown in the analysis of K63-linked triubiquitin (Figure 6). This ability to consider multistate models is in better agreement with the current understanding of solution ensembles than forcing a single representative fit.

Although this AIC protocol is functional in practice, it is important to note that the identified states are not necessarily representative of the full ensemble. Because of the inherently low resolution of SAXS data, increasing subset sizes may be routinely dismissed as long as the average shape and size of the ensemble is modeled by a combination of fewer members. In the future, the addition of further structural information, such as from FRET or NMR experiments, could be incorporated into the Bayesian likelihood function, and this higher resolution data may better differentiate between essential and trivial additional states.[67–71]

This study presents, to our knowledge, the first example of using aMD simulations to rapidly produce atomistic models of experimental SAXS data. In each ubiquitin trimer, the aMD trajectories produced models equal to or better than their cMD counterparts in up to a magnitude less simulation time. This is rooted in the fact that aMD typically supplies not only larger maximum variations in structural RMSDs but also reduces the correlation time by significantly lowering energy barriers between states. However, a wide variety of other possible enhancement methods could be employed to the same end.[11,12,17,18,72] Nonetheless, many of these methods may require producing multiple copies of the same system that must all be simulated simultaneously or performing simulations over a period of time to compute a history-dependent bias, potentially creating large computational overhead. In contrast, aMD requires no additional system replicas and is straightforward to apply with minimal computational cost.[73,74] Therefore, aMD may act as a user-friendly method for those interested in atomistic SAXS modeling without access to high-performance computing resources.

Our protocol was applied to simulations of seven ubiquitin trimers with varying linkage types. Regardless of one- or two-state models, both aMD and cMD results agree with previous studies that K63-linked oligomers adopt extended conformations in solution in contrast to the typically compacted state of K48-linked polyubiquitin chains.[75–78] In addition, our results for the so-called atypical chains reveal that K6 and K11 trimers favor more compact states,[79] whereas K29 trimers favor a more open/intermediate state. The biological functions of these atypical chains are not well-characterized, and it has been suggested that there is a built-in redundancy in the recognition and/or signaling of ubiquitin chains.[80] For example, K48-linked ubiquitin chains are the canonical proteasomal degradative signal, but K11-linked ubiquitin chains can also target substrates to the proteasome in the context of cell cycle progression.[81] As another example, K6- and K63-linked chains are both involved in DNA damage response.[82] Lastly, K29-linked chains are thought to have roles in both proteasomal degradation signaling and regulation of mRNA stability.[83,84] Thus, it appears that a compact structure does not always lead to one biological function and an open structure to another; thus, there is currently no simple correlation between oligomer compaction and cellular response.

Specific to the role of polyubiquitin chains as proteasomal degradation signals, it has been previously theorized that the tight compaction of K48-linked chains might allow for the formation of octamers that could bridge the ~90 Å distance from ubiquitin receptors Rpn10 to Rpn13 of the 26S proteasome.[85,86] Interestingly, our two-state model of K11-linked trimers identified a state with the same globular size as that of the K48-linked trimer, suggesting that it is possible for homotypical K11-linked octamers to satisfy a similar spatial ensemble as that of the K48 octamer. However, proteasomes are able to distinguish between these two modes of polyubiquitin linkage.[87] In our K48-linked trimer models, the hydrophobic patch is packaged around the I44 residue of the central and proximal monomers (Figure 8), and the alternative hydrophobic binding site of the I36 patch is exposed.[88] The opposite scenario is true in one state of our K11-linked model, and the fully compact state of K11 trimers buries all three I36 patch residues while simultaneously exposing all three I44 patch residues. Indeed, domains have been observed to bind selectively with these two sites,[89] and their differing levels of exposure likely contribute to the disparate modes of interactions between K11- and K48-linked polyubiquitin chains and the proteasome.

The overall varying degree of flexibility associated with each different linkage is also likely the major contributing factor to their biological roles.[80,87,90,91] The dynamic nature of ubiquitin chains that can allow their recognition by numerous protein partners is perhaps dictated by the linkage position. Comparison of our models of native isopeptide and non-native thiolene linkages suggest that certain positions (i.e., K48) may be significantly more affected by the local chemistry of the linkage than others (i.e., K63). Some of these effects may be more apparent in trimers than dimers,[35] and these differences may propagate in longer polyubiquitin chains into effects that are significant enough to be discerned by the relatively low-resolution SAXS observations. The methodology developed in this study is primed for investigating longer ubiquitin chains as well as more complex ubiquitin systems including ubiquitin chains bound to receptor proteins, mixed chains, and branched chains, among others, in an effort to better understand the complex role of ubiquitin chain function in cells.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Dr. Srinivas Chakravarthy at Argonne National Laboratory for assistance during data collection and interpretation. Work in the Wereszczynski group was supported by the National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health under award numbers R15GM114758 and R35GM119647. Work in the Strieter lab was supported by NIGMS (R01GM110543). SAXS experiments were performed at BioCAT, which is located at the Advanced Photon Source, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357, and was supported by NIGMS grant P41GM103622. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work used the Extreme Science and Engineering Discovery Environment (XSEDE),[92] which is supported by National Science Foundation Grant ACI-1053575.

References

1. Henzler-Wildman K, Kern D. Dynamic personalities of proteins. Nature. 2007; 450:964–972. [PubMed: 18075575]
2. Boldon L, Laliberte F, Liu L. Review of the fundamental theories behind small angle X-ray scattering, molecular dynamics simulations, and relevant integrated application. Nano Rev. 2015; 6:25661. [PubMed: 25721341]
3. Kachala, M., Valentini, E., Svergun, DI. Intrinsically Disordered Proteins Studied by NMR Spectroscopy. Felli, IC., Pierattelli, R., editors. Springer; 2015. p. 261-289.
4. Tsutakawa SE, Hura GL, Frankel KA, Cooper PK, Tainer JA. Structural analysis of flexible proteins in solution by small angle X-ray scattering combined with crystallography. J Struct Biol. 2007; 158:214–223. [PubMed: 17182256]
5. Fagan RP, Albesa-Jove D, Qazi O, Svergun DI, Brown KA, Fairweather NF. Structural insights into the molecular organization of the S-layer from Clostridium difficile. Mol Microbiol. 2009; 71:1308–1322. [PubMed: 19183279]
6. Gersch M, Famulla K, Dahmen M, Gobl C, Malik I, Richter K, Korotkov VS, Sass P, Rubsamen-Schaeff H, Madl T, Brotz-Oesterhelt H, Sieber SA. AAA+ chaperones and acyldepsipeptides activate the ClpP protease via conformational control. Nat Commun. 2015; 6:6320. [PubMed: 25695750]
7. Cornilescu G, Didychuk AL, Rodgers ML, Michael LA, Burke JE, Montemayor EJ, Hoskins AA, Butcher SE. Structural Analysis of Multi-Helical RNAs by NMR-SAXS/WAXS: Application to the U4/U6 di-snRNA. J Mol Biol. 2016; 428:777–789. [PubMed: 26655855]
8. Kaptein R, Wagner G. NMR studies of membrane proteins. J Biomol NMR. 2015; 61:181–184. [PubMed: 25840906]
9. Henderson R. The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. Q Rev Biophys. 1995; 28:171–193. [PubMed: 7568675]
10. Hamelberg D, Mongan J, McCammon JA. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. J Chem Phys. 2004; 120:11919–11929. [PubMed: 15268227]
11. Hritz J, Oostenbrink C. Hamiltonian replica exchange molecular dynamics using soft-core interactions. J Chem Phys. 2008; 128:144121. [PubMed: 18412437]
12. Nymeyer H, Gnanakaran S, Garcia AE. Atomic simulations of protein folding, using the replica exchange algorithm. Methods Enzymol. 2004; 383:119–149. [PubMed: 15063649]
13. Salomon-Ferrer R, Gotz AW, Poole D, Le Grand S, Walker RC. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. J Chem Theory Comput. 2013; 9:3878–3888. [PubMed: 26592383]

14. Gotz AW, Williamson MJ, Xu D, Poole D, Le Grand S, Walker RC. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. J Chem Theory Comput. 2012; 8:1542–1555. [PubMed: 22582031]
15. Tanner DE, Phillips JC, Schulten K. GPU/CPU Algorithm for Generalized Born/Solvent-Accessible Surface Area Implicit Solvent Calculations. J Chem Theory Comput. 2012; 8:2521–2530. [PubMed: 23049488]
16. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, Lindahl E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX. 2015; 12:19–25.
17. Cuendet MA, Tuckerman ME. Free Energy Reconstruction from Metadynamics or Adiabatic Free Energy Dynamics Simulations. J Chem Theory Comput. 2014; 10:2975–2986. [PubMed: 26588271]
18. Laio A, Parrinello M. Escaping free-energy minima. Proc Natl Acad Sci U S A. 2002; 99:12562–12566. [PubMed: 12271136]
19. Hamelberg D, de Oliveira CA, McCammon JA. Sampling of slow diffusive conformational transitions with accelerated molecular dynamics. J Chem Phys. 2007; 127:155102. [PubMed: 17949218]
20. Allison JR. Using simulation to interpret experimental data in terms of protein conformational ensembles. Curr Opin Struct Biol. 2017; 43:79–87. [PubMed: 27940377]
21. Curtis JE, Raghunandan S, Nanda H, Krueger S. SASSIE: A program to study intrinsically disordered biological molecules and macromolecular ensembles using experimental scattering restraints. Comput Phys Commun. 2012; 183:382–389.
22. Perkins SJ, Wright DW, Zhang H, Brookes EH, Chen J, Irving TC, Krueger S, Barlow DJ, Edler KJ, Scott DJ, Terrill NJ, King SM, Butler PD, Curtis JE. Atomistic modelling of scattering data in the Collaborative Computational Project for Small Angle Scattering (CCP-SAS). J Appl Crystallogr. 2016; 49:1861–1875. [PubMed: 27980506]
23. Howell SC, Qiu X, Curtis JE. Monte Carlo simulation algorithm for B-DNA. J Comput Chem. 2016; 37:2553–2563. [PubMed: 27671358]
24. Cragnell C, Durand D, Cabane B, Skepö M. Coarse-grained modelling of the intrinsically disordered protein Histatin 5 in solution. Monte Carlo simulations in combination with SAXS. Proteins: Struct, Funct, Genet. 2016; 84:777–791. [PubMed: 26914439]
25. Pelikan M, Hura GL, Hammel M. Structure and flexibility within proteins as identified through small angle X-ray scattering. Gen Physiol Biophys. 2009; 28:174–189.
26. Bernado P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI. Structural characterization of flexible proteins using small-angle X-ray scattering. J Am Chem Soc. 2007; 129:5656–5664. [PubMed: 17411046]
27. Tria G, Mertens HD, Kachala M, Svergun DI. Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. IUCrJ. 2015; 2:207–217.
28. Yang S, Blachowicz L, Makowski L, Roux B. Multidomain assembled states of Hck tyrosine kinase in solution. Proc Natl Acad Sci U S A. 2010; 107:15757–15762. [PubMed: 20798061]
29. Hawkins DM. The problem of overfitting. J Chem Inf Comput Sci. 2004; 44:1–12. [PubMed: 14741005]
30. Chau V, Tobias JW, Bachmair A, Marriott D, Ecker DJ, Gonda DK, Varshavsky A. A multiubiquitin chain is confined to specific lysine in a targeted short-lived protein. Science. 1989; 243:1576–1583. [PubMed: 2538923]
31. Nathan JA, Kim HT, Ting L, Gygi SP, Goldberg AL. Why do cellular proteins linked to K63-polyubiquitin chains not associate with proteasomes? EMBO J. 2013; 32:552–565. [PubMed: 23314748]
32. Chen ZJ. Ubiquitination in signaling to and activation of IKK. Immunol Rev. 2012; 246:95–106. [PubMed: 22435549]
33. Valkevich EM, Guenette RG, Sanchez NA, Chen YC, Ge Y, Strieter ER. Forging isopeptide bonds using thiol-ene chemistry: site-specific coupling of ubiquitin molecules for studying the activity of isopeptidases. J Am Chem Soc. 2012; 134:6916–6919. [PubMed: 22497214]

34. Pham GH, Strieter ER. Peeling away the layers of ubiquitin signaling complexities with synthetic ubiquitin-protein conjugates. Curr Opin Chem Biol. 2015; 28:57–65. [PubMed: 26093241]

35. Pham GH, Rana AS, Korkmaz EN, Trang VH, Cui Q, Strieter ER. Comparison of native and non-native ubiquitin oligomers reveals analogous structures and reactivities. Protein Sci. 2016; 25:456–471. [PubMed: 26506216]

36. Vijay-Kumar S, Bugg CE, Cook WJ. Structure of ubiquitin refined at 1.8 A resolution. J Mol Biol. 1987; 194:531–544. [PubMed: 3041007]

37. Mathew E, Mirza A, Menhart N. Liquid-chromatography-coupled SAXS for accurate sizing of aggregating proteins. J Synchrotron Radiat. 2004; 11:314–318. [PubMed: 15211037]

38. Konarev P, Volkov V, Sokolova A, Koch M, Svergun D. PRIMUS: a Windows PC-based system for small-angle scattering data analysis. J Appl Crystallogr. 2003; 36:1277–1282.

39. Virtanen JJ, Makowski L, Sosnick TR, Freed KF. Modeling the hydration layer around proteins: applications to small-and wide-angle x-ray scattering. Biophys J. 2011; 101:2061–2069. [PubMed: 22004761]

40. Matsumoto ML, Wickliffe KE, Dong KC, Yu C, Bosanac I, Bustos D, Phu L, Kirkpatrick DS, Hymowitz SG, Rape M, Kelley RF, Dixit VM. K11-linked polyubiquitination in cell cycle control revealed by a K11 linkage-specific antibody. Mol Cell. 2010; 39:477–484. [PubMed: 20655260]

41. Virdee S, Ye Y, Nguyen DP, Komander D, Chin JW. Engineered diubiquitin synthesis reveals Lys29-isopeptide specificity of an OTU deubiquitinase. Nat Chem Biol. 2010; 6:750–757. [PubMed: 20802491]

42. Kristariyanto YA, Abdul Rehman SA, Campbell DG, Morrice NA, Johnson C, Toth R, Kulathu Y. K29-selective ubiquitin binding domain reveals structural basis of specificity and heterotypic nature of k29 polyubiquitin. Mol Cell. 2015; 58:83–94. [PubMed: 25752573]

43. Satoh T, Sakata E, Yamamoto S, Yamaguchi Y, Sumiyoshi A, Wakatsuki S, Kato K. Crystal structure of cyclic Lys48-linked tetraubiquitin. Biochem Biophys Res Commun. 2010; 400:329–333. [PubMed: 20728431]

44. Datta AB, Hura GL, Wolberger C. The structure and conformation of Lys63-linked tetraubiquitin. J Mol Biol. 2009; 392:1117–1124. [PubMed: 19664638]

45. Frisch, MJ., Trucks, GW., Schlegel, HB., Scuseria, GE., Robb, MA., Cheeseman, JR., Scalmani, G., Barone, V., Mennucci, B., Petersson, GA., Nakatsuji, H., Caricato, M., Li, X., Hratchian, HP., Izmaylov, AF., Bloino, J., Zheng, G., Sonnenberg, JL., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Vreven, T., Montgomery, JA., Jr, Peralta, JE., Ogliaro, F., Bearpark, M., Heyd, JJ., Brothers, E., Kudin, KN., Staroverov, VN., Kobayashi, R., Normand, J., Raghavachari, K., Rendell, A., Burant, JC., Iyengar, SS., Tomasi, J., Cossi, M., Rega, N., Millam, JM., Klene, M., Knox, JE., Cross, JB., Bakken, V., Adamo, C., Jaramillo, J., Gomperts, R., Stratmann, RE., Yazyev, O., Austin, AJ., Cammi, R., Pomelli, C., Ochterski, JW., Martin, RL., Morokuma, K., Zakrzewski, VG., Voth, GA., Salvador, P., Dannenberg, JJ., Dapprich, S., Daniels, AD., Farkas, O., Foresman, JB., Ortiz, JV., Cioslowski, J., Fox, DJ. Gaussian 09. Gaussian Inc; Wallingford, CT: 2009. revision D.01

46. Jorgensen W, Chandrasekhar J, Madura J, Impey R, Klein M. Comparison of simple potential functions for simulating liquid water. J Chem Phys. 1983; 79:926–935.

47. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. J Chem Theory Comput. 2015; 11:3696–3713. [PubMed: 26574453]

48. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. J Chem Phys. 1984; 81:3684–3690.

49. Krautler V, Van Gunsteren WF, Hunenberger PH. A fast SHAKE: Algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. J Comput Chem. 2001; 22:501–508.

50. Darden T, York D, Pedersen L. Particle mesh Ewald - an N log(N) method for Ewald sums in large systems. J Chem Phys. 1993; 98:10089–10092.

51. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM, Onufriev A, Simmerling C, Wang B, Woods RJ. The Amber biomolecular simulation programs. J Comput Chem. 2005; 26:1668–1688. [PubMed: 16200636]

52. Roe DR, Cheatham TE. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. J Chem Theory Comput. 2013; 9:3084–3095. [PubMed: 26583988]

53. Wereszczynski, J., McCammon, JA. Computational Drug Discovery and Design. Baron, R., editor. Humana Press; 2012. p. 515-524.

54. Shao J, Tanner SW, Thompson N, Cheatham TE. Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. J Chem Theory Comput. 2007; 3:2312–2334. [PubMed: 26636222]

55. Torda AE, van Gunsteren WF. Algorithms for clustering molecular dynamics configurations. J Comput Chem. 1994; 15:1331–1340.

56. Svergun D, Barberato C, Koch MHJ. *CRYSOL* – a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. J Appl Crystallogr. 1995; 28:768–773.

57. Moore PB. Small-angle scattering Information content and error analysis. J Appl Crystallogr. 1980; 13:168–175.

58. Schneidman-Duhovny D, Hammel M, Tainer JA, Sali A. Accurate SAXS profile computation and its assessment by contrast variation experiments. Biophys J. 2013; 105:962–974. [PubMed: 23972848]

59. Ravikumar KM, Huang W, Yang S. Fast-SAXS-pro: a unified approach to computing SAXS profiles of DNA, RNA, protein, and their complexes. J Chem Phys. 2013; 138:024112. [PubMed: 23320673]

60. Köfinger J, Hummer G. Atomic-resolution structural information from scattering experiments on macromolecules in solution. Phys Rev E Stat Nonlin Soft Matter Phys. 2013; 87:052712. [PubMed: 23767571]

61. Hines KE. A primer on Bayesian inference for biophysical systems. Biophys J. 2015; 108:2103–2113. [PubMed: 25954869]

62. Putnam CD, Hammel M, Hura GL, Tainer JA. X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. Q Rev Biophys. 2007; 40:191–285. [PubMed: 18078545]

63. Rambo RP, Tainer JA. Accurate assessment of mass, models and resolution by small-angle scattering. Nature. 2013; 496:477–481. [PubMed: 23619693]

64. Spill YG, Kim SJ, Schneidman-Duhovny D, Russel D, Webb B, Sali A, Nilges M. SAXS Merge: an automated statistical method to merge SAXS profiles using Gaussian processes. J Synchrotron Radiat. 2014; 21:203–208. [PubMed: 24365937]

65. Akaike, H. International Encyclopedia of Statistical Science. Lovric, M., editor. Springer; Berlin, Heidelberg: 2011. p. 25-25.

66. Konarev PV, Svergun DI. A posteriori determination of the useful data range for small-angle scattering experiments on dilute monodisperse systems. IUCrJ. 2015; 2:352–360.

67. Fisher CK, Huang A, Stultz CM. Modeling intrinsically disordered proteins with bayesian statistics. J Am Chem Soc. 2010; 132:14919–14927. [PubMed: 20925316]

68. Gurry T, Ullman O, Fisher CK, Perovic I, Pochapsky T, Stultz CM. The dynamic structure of alpha-synuclein multimers. J Am Chem Soc. 2013; 135:3865–3872. [PubMed: 23398399]

69. van de Meent J-W, Bronson JE, Wiggins CH, Gonzalez RL. Empirical Bayes methods enable advanced population-level analyses of single-molecule FRET experiments. Biophys J. 2014; 106:1327–1337. [PubMed: 24655508]

70. Murphy RR, Danezis G, Horrocks MH, Jackson SE, Klenerman D. Bayesian Inference of Accurate Population Sizes and FRET Efficiencies from Single Diffusing Biomolecules. Anal Chem. 2014; 86:8603–8612. [PubMed: 25105347]

71. Bonomi M, Pellarin R, Kim SJ, Russel D, Sundin BA, Riffle M, Jaschob D, Ramsden R, Davis TN, Muller EG, Sali A. Determining protein complex structures based on a Bayesian model of in vivo Förster resonance energy transfer (FRET) data. Mol Cell Proteomics. 2014; 13:2812–2823. [PubMed: 25139910]

72. Chen PC, Hub JS. Interpretation of solution x-ray scattering by explicit-solvent molecular dynamics. Biophys J. 2015; 108:2573–2584. [PubMed: 25992735]

73. Pierce LC, Salomon-Ferrer R, Augusto F, de Oliveira C, McCammon JA, Walker RC. Routine access to millisecond time scale events with accelerated molecular dynamics. J Chem Theory Comput. 2012; 8:2997–3002. [PubMed: 22984356]

74. Wang Y, Harrison CB, Schulten K, McCammon JA. Implementation of accelerated molecular dynamics in NAMD. Comput Sci Discovery. 2011; 4:015002.

75. Liu Z, Gong Z, Jiang W-X, Yang J, Zhu W-K, Guo D-C, Zhang W-P, Liu M-L, Tang C. Lys63-linked ubiquitin chain adopts multiple conformational states for specific target recognition. eLife. 2015; 4:e05767.

76. Ye Y, Blaser G, Horrocks MH, Ruedas-Rama MJ, Ibrahim S, Zhukov AA, Orte A, Klenerman D, Jackson SE, Komander D. Ubiquitin chain conformation regulates recognition and activity of interacting proteins. Nature. 2012; 492:266–270. [PubMed: 23201676]

77. Tenno T, Fujiwara K, Tochio H, Iwai K, Morita EH, Hayashi H, Murata S, Hiroaki H, Sato M, Tanaka K, Shirakawa M. Structural basis for distinct roles of Lys63- and Lys48-linked polyubiquitin chains. Genes Cells. 2004; 9:865–875. [PubMed: 15461659]

78. Varadan R, Assfalg M, Haririnia A, Raasi S, Pickart C, Fushman D. Solution conformation of Lys63-linked di-ubiquitin chain provides clues to functional diversity of polyubiquitin signaling. J Biol Chem. 2003; 279:7055–7063. [PubMed: 14645257]

79. Nakasone MA, Livnat-Levanon N, Glickman MH, Cohen RE, Fushman D. Mixed-linkage ubiquitin chains send mixed messages. Structure. 2013; 21:727–740. [PubMed: 23562397]

80. Xu P, Duong DM, Seyfried NT, Cheng D, Xie Y, Robert J, Rush J, Hochstrasser M, Finley D, Peng J. Quantitative Proteomics Reveals the Function of Unconventional Ubiquitin Chains in Proteasomal Degradation. Cell. 2009; 137:133–145. [PubMed: 19345192]

81. Min M, Mevissen TE, De Luca M, Komander D, Lindon C. Efficient APC/C substrate degradation in cells undergoing mitotic exit depends on K11 ubiquitin linkages. Mol Biol Cell. 2015; 26:4325–4332. [PubMed: 26446837]

82. Kulathu Y, Komander D. Atypical ubiquitylation - the unexplored world of polyubiquitin beyond Lys48 and Lys63 linkages. Nat Rev Mol Cell Biol. 2012; 13:508–523. [PubMed: 22820888]

83. Wagner SA, Beli P, Weinert BT, Nielsen ML, Cox J, Mann M, Choudhary C. A proteome-wide, quantitative survey of in vivo ubiquitylation sites reveals widespread regulatory roles. Mol Cell Proteomics. 2011; 10:M111.013284.

84. Zhou HL, Geng C, Luo G, Lou H. The p97-UBXD8 complex destabilizes mRNA by promoting release of ubiquitinated HuR from mRNP. Genes Dev. 2013; 27:1046–1058. [PubMed: 23618873]

85. Schreiber A, Peter M. Substrate recognition in selective autophagy and the ubiquitin-proteasome system. Biochim Biophys Acta, Mol Cell Res. 2014; 1843:163–181.

86. Lasker K, Forster F, Bohn S, Walzthoeni T, Villa E, Unverdorben P, Beck F, Aebersold R, Sali A, Baumeister W. Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach. Proc Natl Acad Sci U S A. 2012; 109:1380–1387. [PubMed: 22307589]

87. Grice GL, Nathan JA. The recognition of ubiquitinated proteins by the proteasome. Cell Mol Life Sci. 2016; 73:3497–3506. [PubMed: 27137187]

88. Winget JM, Mayor T. The diversity of ubiquitin recognition: hot spots and varied specificity. Mol Cell. 2010; 38:627–635. [PubMed: 20541996]

89. Reyes-Turcu FE, Horton JR, Mullally JE, Heroux A, Cheng X, Wilkinson KD. The ubiquitin binding domain ZnF UBP recognizes the C-terminal diglycine motif of unanchored ubiquitin. Cell. 2006; 124:1197–1208. [PubMed: 16564012]

90. Pickart CM, Fushman D. Polyubiquitin chains: polymeric protein signals. Curr Opin Chem Biol. 2004; 8:610–616. [PubMed: 15556404]

91. Yau R, Rape M. The increasing complexity of the ubiquitin code. Nat Cell Biol. 2016; 18:579–586. [PubMed: 27230526]

92. Towns J, Cockerill T, Dahan M, Foster I, Gaither K, Grimshaw A, Hazlewood V, Lathrop S, Lifka D, Peterson GD, Roskies R, Scott JR, Wilkens-Diehr N. XSEDE: Accelerating Scientific Discovery. Comput Sci Eng. 2014; 16:62–74.

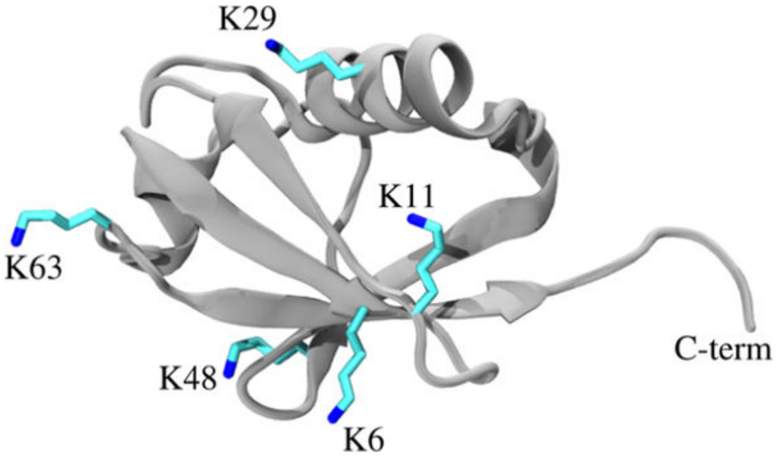


Figure 1.

Crystal structure of ubiquitin monomer (PDB: 1UBQ).[36] Shown in sticks are the five lysine sites presented in this study. Other potential linkage sites exist but are not shown (M1, K27, and K33). The C-terminus (“C-term”) of one ubiquitin monomer forms an isopeptide or thiolene linkage with the amino group at one of these sites.

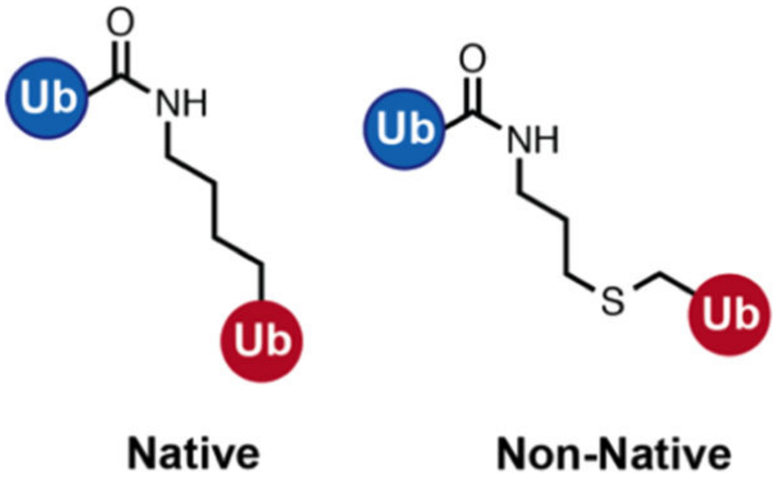


Figure 2.

Depiction of the native isopeptide linkage between ubiquitin monomers (left) and the non-native thiolene linkage used in this study (right).

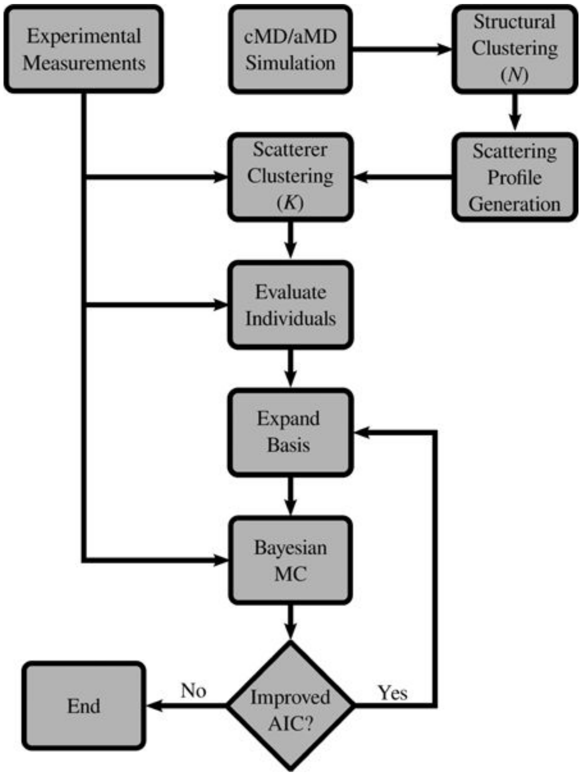

Experimental Measurements
cMD/aMD Simulation
Structural Clustering (*N*)
Scatterer Clustering (*K*)
Scattering Profile Generation
Evaluate Individuals
Expand Basis
Bayesian MC
End
No
Improved AIC?
Yes

Figure 3.

Schematic of the iterative Bayesian ensemble refinement workflow. MD trajectories were first clustered into similar structures that were subsequently clustered based on their scattering patterns. The full permutation of iteratively increasing subset sizes were then used to produce models through a Bayesian Monte Carlo until overfitting was observed.

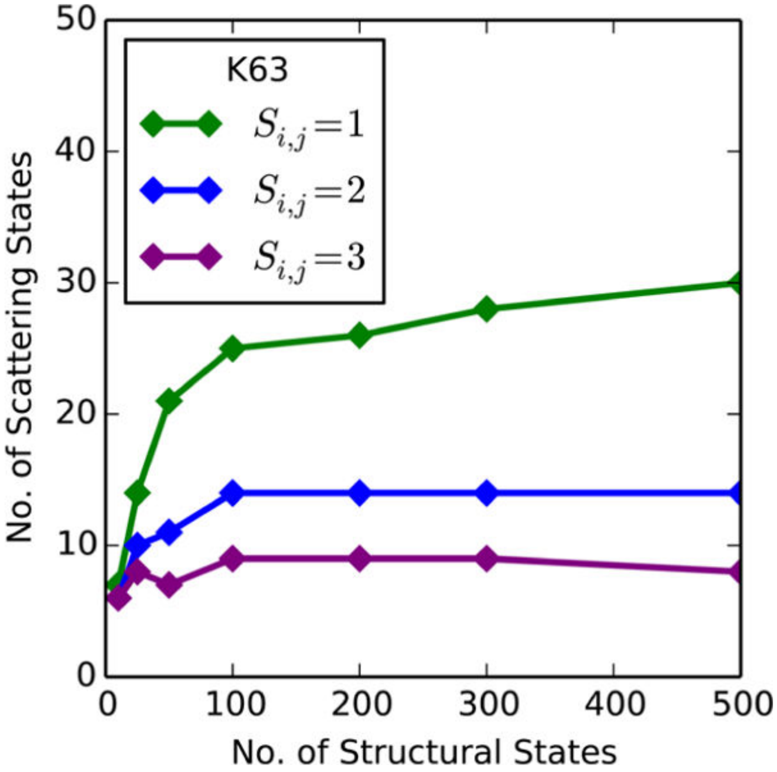


Figure 4.

Number of identified scattering states vs the number of initial structural clusters from the K63 triubiquitin aMD simulation. At a similarity restriction of $S_{i,j}$ = 2, the number of unique scattering states is unaltered when considering more than 100 structural states.

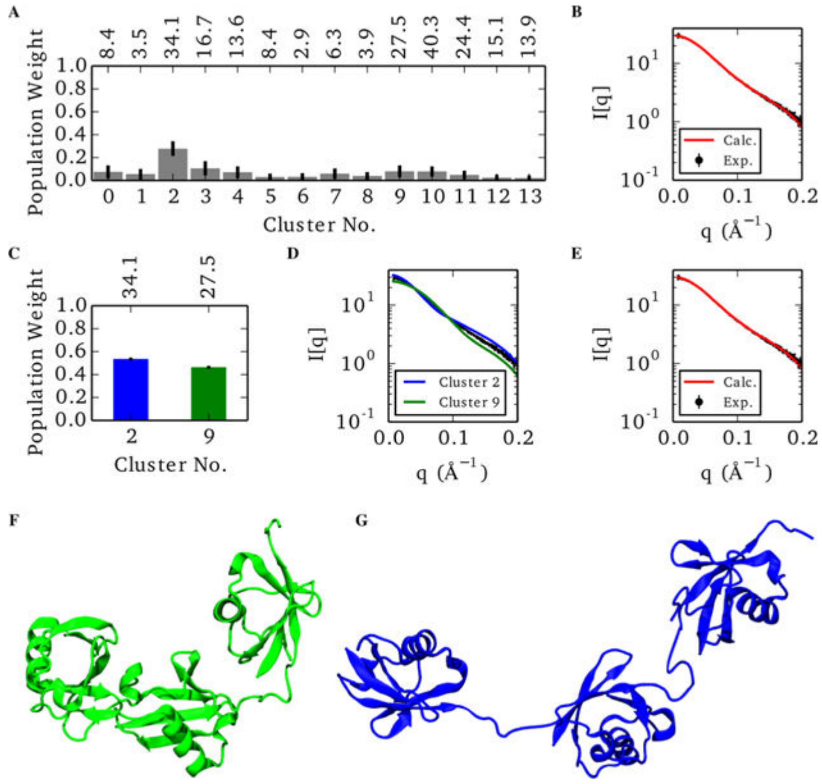


Figure 5.

Representative ensemble of the non-native K63 aMD model using the full 14-member scattering basis (A,B) and the best two-state model (C-G). (A) Populations of the individual states of the 14-member model, where error bars indicate the standard deviation of each marginal posterior distribution. Individual goodness-of-fit values are denoted above the corresponding population bar. (B) Comparison of the ensemble averaged scattering profile (red) against the experimental data (black). (C) Relative populations of the best two-state combination with individual χ^2_{free} values above the plot. (D) Individual scattering profiles of each state. (E) Ensemble-averaged scattering of the two-state model compared with experimental data. (F) Representative member of the compacted cluster. (G) Representative member of the extended cluster.

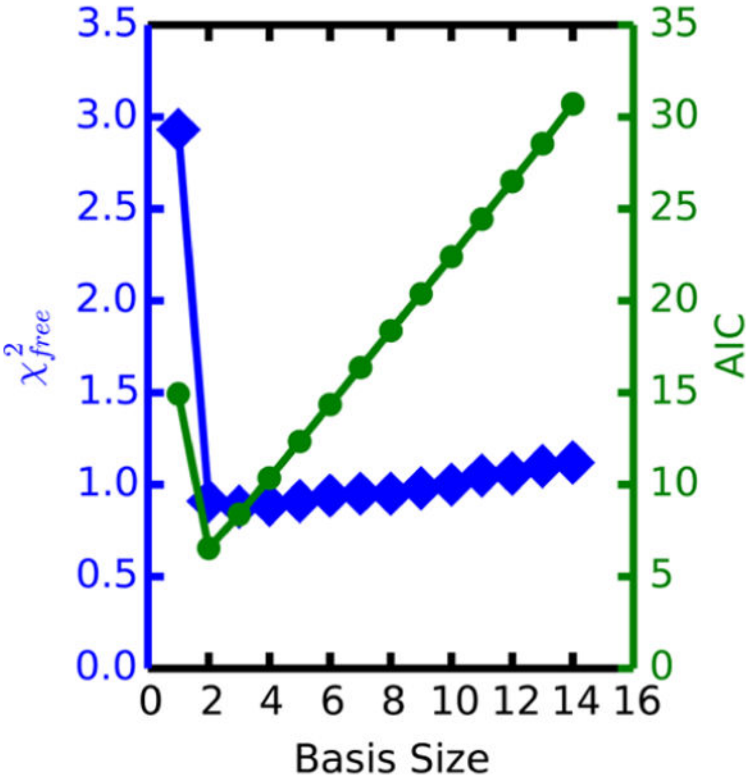


Figure 6.

Lowest observed reduced χ^2_{free} goodness-of-fit (blue) and the corresponding AIC value (green) for each basis subset size in the non-native K63 aMD model. Although the χ^2_{free} value of a 3-state ensemble is a modest improvement over the 2-state model, the increased AIC value suggests that this is a result of increased overfitting by the 3-state case.

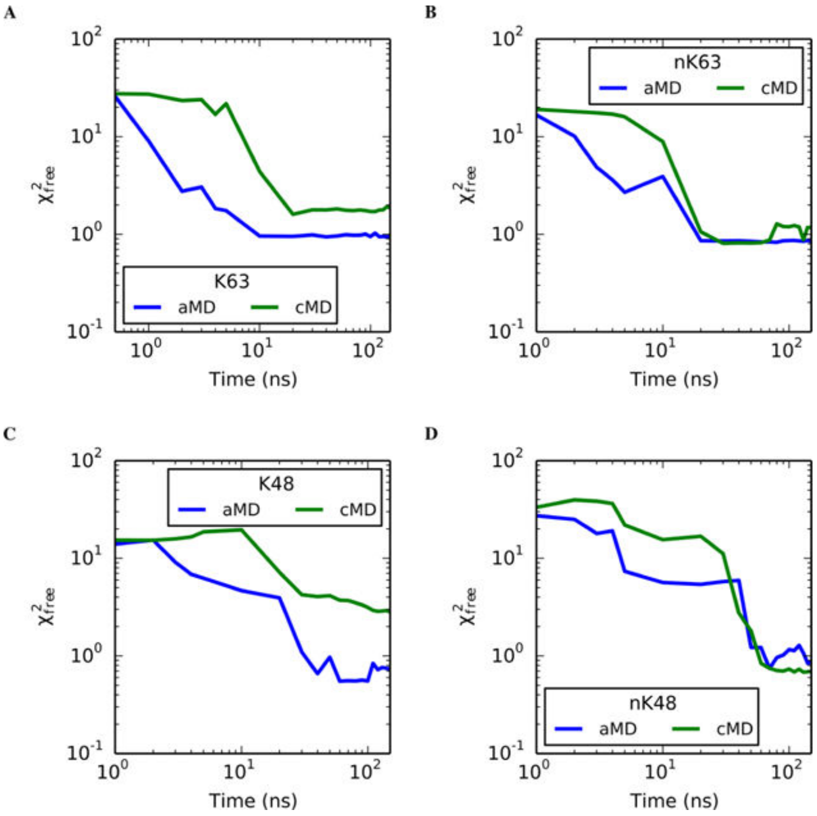


Figure 7.

Reduced χ^2_{free} goodness-of-fit of identified ensembles vs sampling time for aMD (blue) and cMD (green) simulations of (a) non-native K63, (b) native K63, (c) non-native K48, and (d) native K48 linkages. Remaining systems can be found in Figure S11. In the systems presented above, the aMD simulations are the quickest to escape poor initial models, as shown by the more rapid initial decrease in χ^2_{free} in comparison to the cMD models.

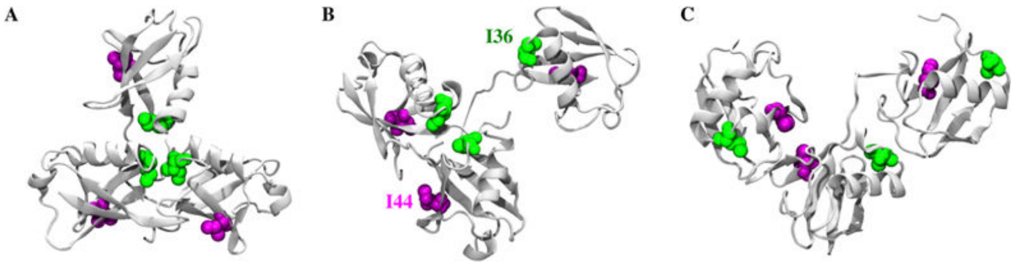


Figure 8.

Orientations of I36 (green) and I44 (purple) hydrophobic residues in the (a) closed K11-trimer state, (b) open K11-trimer state, and (c) K48-linked trimer model. Although both the open K11 and K48 configurations display similar levels of compaction according to R_g, they expose different access to the hydrophobic patch: I44 in K11 and I36 in K48. Furthermore, the highly compacted K11 state buries all three I36 moieties into the same region, thus exposing all three I44 moieties. The deviations in modes of hydrophobic exposure may contribute to the binding disparities of K11- and K48-linked polyubiquitin chains with both proteasomes and deubiquitinating enzymes.

Table 1

R_g for Each System As Determined from a Guinier Analysis of the Experimental Data, Analysis of the cMD and aMD Trajectories, and from the Bayesian Ensemble Fitting Protocol[a]

system	experiment (Å)	cMD (Å)	aMD (Å)	ensemble (Å)
K6	22.9 ± 0.1	19.0 ± 0.1	19.3 ± 0.2	21.8 ± 0.1
K11	21.4 ± 0.1	20.5 ± 0.1	19.5 ± 0.1	21.7 ± 0.6
K29	23.3 ± 0.3	24.6 ± 0.1	20.3 ± 0.2	24.9 ± 1.5
K48	22.3 ± 0.1	23.6 ± 0.3	21.3 ± 0.3	22.4 ± 0.3
nK48	23.7 ± 0.1	22.4 ± 1.0	23.4 ± 0.2	24.1 ± 0.6
K63	28.0 ± 0.1	25.0 ± 0.2	22.5 ± 0.2	28.3 ± 0.3
nK63	27.0 ± 0.2	25.6 ± 0.3	21.1 ± 0.5	28.6 ± 1.1

[a] The R_g of a ubiquitin trimer appears to be directly related to the geometry of the linkage. Furthermore, ensemble reweighting produces better agreement with experimental values than the raw MD trajectories. nK48 and nK63 denote trimers with the native isopeptide linkage.

Table 2

Summary of the aMD Models for Each System[a]

system	weight	R_g (Å)	sep. dist. (Å)	sep. angle (deg)	χ^2_{free}
K6	1.00	21.8 ± 0.1	33.4 ± 1.0	79.8 ± 4.8	2.4
K11	0.47	20.7 ± 0.2	27.4 ± 1.1	61.9 ± 3.6	1.8
	0.53	22.5 ± 0.6	35.9 ± 2.9	78.4 ± 8.2	1.8
	1.00	21.7 ± 0.6	31.9 ± 3.1	70.6 ± 8.9	0.7
K29	1.00	24.9 ± 1.5	44.6 ± 5.8	93.4 ± 12.1	0.9
K48	1.00	22.4 ± 0.3	37.1 ± 1.8	86.1 ± 8.9	0.7
nK48	0.11	29.8 ± 0.5	62.7 ± 1.8	130.9 ± 6.0	39.1
	0.89	23.4 ± 0.2	43.3 ± 0.8	111.1 ± 4.2	1.4
	1.00	24.1 ± 0.6	45.3 ± 2.0	113.2 ± 7.3	0.9
K63	0.54	32.6 ± 0.2	68.9 ± 0.4	120.4 ± 0.1	34.1
	0.46	23.3 ± 0.3	41.0 ± 1.5	97.0 ± 8.3	27.5
	1.00	28.3 ± 0.3	56.0 ± 1.6	109.6 ± 8.3	1.0
nK63	0.46	31.7 ± 1.0	69.8 ± 2.8	149.3 ± 8.3	3.5
	0.54	25.8 ± 0.5	51.2 ± 1.9	121.4 ± 4.8	2.3
	1.00	28.6 ± 1.1	59.8 ± 3.4	134.4 ± 9.7	0.9

[a] Separation distances and angles are measured between the centers of masses of the distal groups and using the center of mass of the central member as the vertex. For the K48 and K63 systems, "nK48" and "nK63" denote the trimers with the native isopeptide linkage. cMD models can be found in Table S2.