Supporting Information for

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, IL, 60616, Department of Chemistry, University of Massachusetts-Amherst, Amherst, MA 01003, Department of Chemistry, University of Wisconsin-Madison, Madison, WI 53706, and Department of Biochemistry and Molecular Biology, University of Massachusetts-Amherst, Amherst, MA 01003

E-mail: estrieter@chem.umass.edu; jwereszc@iit.edu

[*]To whom correspondence should be addressed

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

[‡]Department of Chemistry, University of Massachusetts-Amherst, Amherst, MA 01003

[¶]Department of Chemistry, University of Wisconsin-Madison, Madison, WI 53706

[§]Department of Biochemistry and Molecular Biology, University of Massachusetts-Amherst, Amherst, MA 01003

[‖]Contributed equally to this work

S1 AIC Example

When fitting an ensemble of models to low-dimensional data, it is important that the ensemble only considers the minimum number of required parameters and that the model does not achieve a high quality of fit through the addition of unwarranted populations. In the main body of this study, we use the Akaike Information Criterion (AIC) to identify the point at which an unwarranted scattering state is added to the ensemble,[1] but here we present a simplified scenario in which we model the function $f(x) = x^2 + x + \delta$ over $x = 0$ to 1, where δ represents random Gaussian noise, using a variety of fitting functions (Figure S1). If we fit the data with a simple linear regression ($f(x) = Ax + B$), then we yield a model with a reduced χ^2 of 12.03 and an AIC value of 244.52. If we instead choose to model the data with a lone parabolic function ($f(x) = Ax^2$), then the fit yields a reduced χ^2 of 4.42 and an AIC value of 90. The subsequent drop in AIC comes not only from the improvement of χ^2, but also because fewer parameters are used in the fitting process ($f(x) = Ax^2$ has a single parameter, A, while $f(x) = Ax + B$ has two parameters, A and B). Fitting according to the underlying function ($f(x) = Ax^2 + Bx$) yields a model with a reduced χ^2 of 0.44 and an AIC value of 12.77, which shows that improvement of χ^2 from the addition of a linear term to the parabolic function is justified by a drastic improvement in AIC ($\Delta AIC \sim 88$). Lastly, we consider a cubic spline of the noisy data, which models the observed data points with a reduced χ^2 of 0.25, nearly a factor of 2 better than the linear combination. However, the

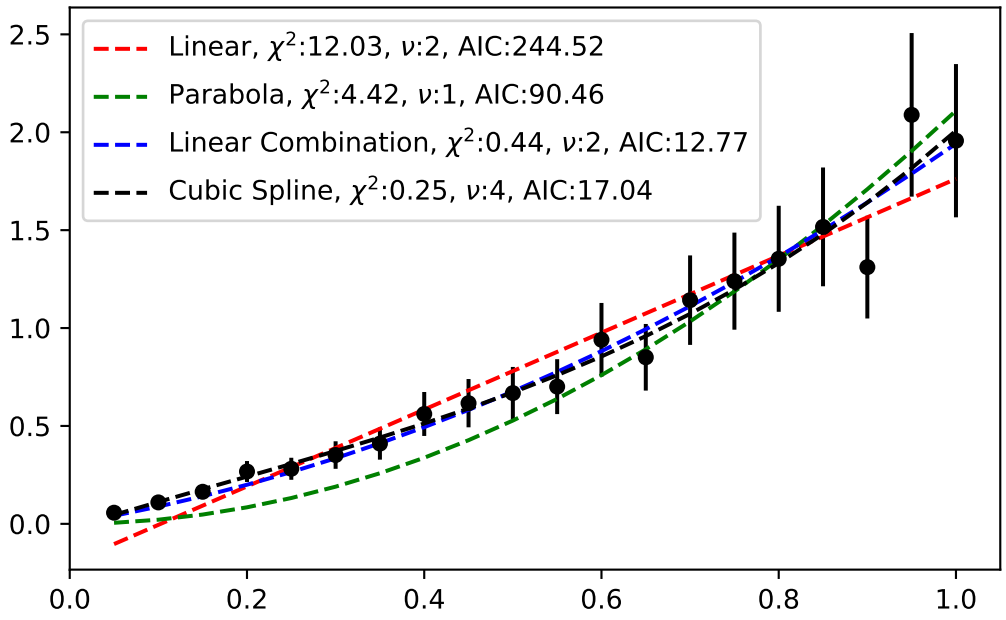

Figure S1: Test case for AIC using the function $f(x) = x^2 + x$ with Gaussian noise (black dots) fit by a linear function (red), a parabolic function (green), $f(x) = Ax^2 + Bx$ (blue), and a cubic spline (black). The figure legend includes the reduced χ^2 goodness-of-fit, the number of model parameters (ν), and the resulting AIC value. The AIC parameter is able to distinguish the linear combination as the most appropriate model instead of the cubic spline, which is the best model according to χ^2 value alone.

AIC actually increases by 5, signifying the addition of unwarranted parameters. This trend is expected since we know that the true function is $f(x) = x^2 + x$. Thus, we observe that the AIC metric is able to distinguish models that improve goodness-of-fit to the underlying data (the linear combination) from those that improve goodness-of-fit by the addition of extraneous parameters that fit to the noise (the cubic spline).

S2 Elaboration of χ^2_{free}

The colloquial measure for model quality is described by the χ^2 metric:

$$\chi^2 = \sum_q \frac{(I_t(q) - I_e(q))^2}{\sigma(q)^2} \tag{S1}$$

where $I_t(q)$ is the theoretical intensity of the model, $I_e(q)$ is the experimental scattering amplitude, and $\sigma^2(q)$ is the experimentally observed error in intensity. Searching for the model with minimum χ^2 is synonymous with maximizing the likelihood function. Because the interpretation of χ^2 is largely dependent on the number of data points, model quality is typically reported as the *reduced* χ^2 value ($\chi^2_{red} = \chi^2/N$), with a $\chi^2_{red} \approx 1$ considered to be a good fit.

One underlying assumption to χ^2-based metrics is that the errors are uncorrelated and Gaussian. This is troublesome for a traditional SAXS experiment, where data points are largely over-sampled and highly correlated with neighboring measurements in q. As a result, the standard χ^2 metric may routinely over-fit SAXS models due to overestimated degrees of freedom. This has led to the recent development of the χ^2_{free} metric, which is based on the Nyquist-Shannon sampling limit of SAXS measurement:[2]

$$\chi^2_{free} = \sum_{q \epsilon S} \frac{(I_t(q) - I_e(q))^2}{\sigma^2(q)} \tag{S2}$$

The sum is carried out for only for a single q value in each Shannon Channel, S, where the number of channels is determined by particle size and data quality ($N_s = q_{max} \cdot D_{max}/\pi$). To ensure that an adequate combination of q points are considered, the reported χ^2_{free} value is the median of over 2,000 random samplings. Analogously to χ^2, a reduced χ^2_{free} can be calculated that is normalized according to the number of channels instead of total number of q points. It has been previously shown that the use of χ^2_{free} in place of χ^2 provides more accurate assessments of model quality and is less prone to over-fitting.[2]

S3 Note on system setup

In an ideal case, simulations of all seven ubiquitin systems would be initiated from x-ray crystal structures since these represent a more physical state than *de novo* models constructed manually. Because of this, tetra-ubiquitin crystal structures were used when possible; however, no tri or tetra-ubiquitin structures exist for the majority of systems studied here. At-

tempts were made to construct all remaining systems by overlapping the central monomer of the corresponding di-ubiquitin crystal structures. In the case of the K11 linked system, this approach led to significant steric clash between the distal ubiquitin monomers so the trimer was instead constructed by adding one ubiquitin monomer to the dimer crystal structure.

S4 Selection of aMD Parameters

Accelerated molecular dynamics (aMD) enhances simulation sampling by reducing potential barriers between states by the addition of a boost potential.[3] This boost potential is of the form:

$$\Delta V(x) = \begin{cases} 0 & \text{if } V(x) \geq E_p \\ \frac{(E_p - V(x))^2}{\alpha + E_p - V(x)} & \text{if } V(x) < E_p \end{cases} \tag{S3}$$

where $V(x)$ is the potential energy when the system has configuration x, E_p is the threshold energy value for applying the boost, and α is the acceleration factor. Selecting the values of α and E_p is non-deterministic, but the standard procedure is to conduct a short conventional MD (cMD) simulation and then calculate α and E_p from the average energy of the cMD simulation using the following relations:

$$\begin{cases} E_p = \langle V_p \rangle + \frac{1}{5} n_{atoms} \\ \alpha_p = \frac{1}{5} n_{atoms} \end{cases} \tag{S4}$$

where $\langle V_p \rangle$ is the average potential energy of the cMD simulation, and n_{atoms} is the number of explicit atoms in the system. Furthermore, additional sampling benefits can be achieved by applying a separate boost to the dihedral component of the energy landscape:

$$\begin{cases} E_d = \langle V_d \rangle + 4 n_{residues} \\ \alpha_d = \frac{4}{5} n_{residues} \end{cases} \tag{S5}$$

where E_d is the dihedral energy threshold, α_d is the dihedral acceleration factor, $\langle V_d \rangle$ is the average dihedral energy from the cMD simulation, and $n_{residues}$ is the number of solute residues. It should be stressed that these values represent initial approximations and may not provide adequate sampling increases upon inspection of the trajectories. In this case, it may be necessary to select more aggressive values for either the energy threshold or the accelerating factor, or in some cases both. For the case of the K6-linked trimer, improved sampling was not achieved by the "standard" protocol, but improved sampling was observed by incrementing E_p and E_d by α_p and α_d, respectively.

References

(1) Akaike, H. In *International Encyclopedia of Statistical Science*; Lovric, M., Ed.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2011; pp 25–25.

(2) Rambo, R.P.; Tainer, J.A. Accurate assessment of mass, models and resolution by small-angle scattering. *Nature* **2013**, *496*, 477–481.

(3) Wereszczynski, J.; McCammon, J.A. In *Computational Drug Discovery and Design*; Baron, R., Ed.; Humana Press, 2012; pp 515–524.

Table S1: PDB IDs of the structures used to build the seven tri-ubiquitin systems, along with the values of the aMD variables used. Note that for the K48 and K6 systems, E_p and E_d were each incremented by the value of the corresponding α in order to attain the desired sampling level. All energies given in units of kcal/mol.

System	PDB used	E_p	α_p	E_d	α_d
K11	3NOB + 1UBQ	-159,213	10,551	3,816	182
K29	2 x 4S22	-253,208	13,270	3,802	182
K48	2 x 3ALB	-121,485	8,690	3,982	182
nK48	2 x 3ALB	-130,207	8,690	3,796	182
K6	2 x 2XK5	-106,162	7,649	3,992	182
K63	3HM3	-229,014	15,051	3,806	182
nK63	3HM3	-232,019	12,040	3,809	182

Table S2: Summary of the cMD models for each system. 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.

	Weights	R_g (Å)	Sep. Dist. (Å)	Sep. Angle (°)	χ^2_{free}
K6 - cMD	1.00	20.4±0.2	27.7±1.3	67.0±3.7	8.7
K11 - cMD	1.00	21.5±0.4	34.3±2.1	84.3±7.2	0.8
K29 - cMD	1.00	25.6±0.8	50.2±3.1	115.4±12.6	0.9
K48 - cMD	1.00	24.3±0.6	47.2±2.4	121.7±8.0	2.8
nK48 - cMD	0.56	22.7±0.2	35.8±0.5	76.5±2.0	6.2
	0.44	26.9±0.3	55.2±1.1	131.3±3.4	11.9
	1.00	24.5±0.4	44.4±1.2	100.8±3.9	0.7
K63 - cMD	0.43	31.4±0.4	64.7±1.3	113.7±2.9	28.3
	0.57	25.1±0.4	49.6±1.4	122.2±3.3	11.7
	1.00	27.8±0.6	56.2±1.9	118.5±4.4	1.9
nK63 - cMD	1.00	27.0±0.6	54.6±2.1	123.2±6.6	1.2

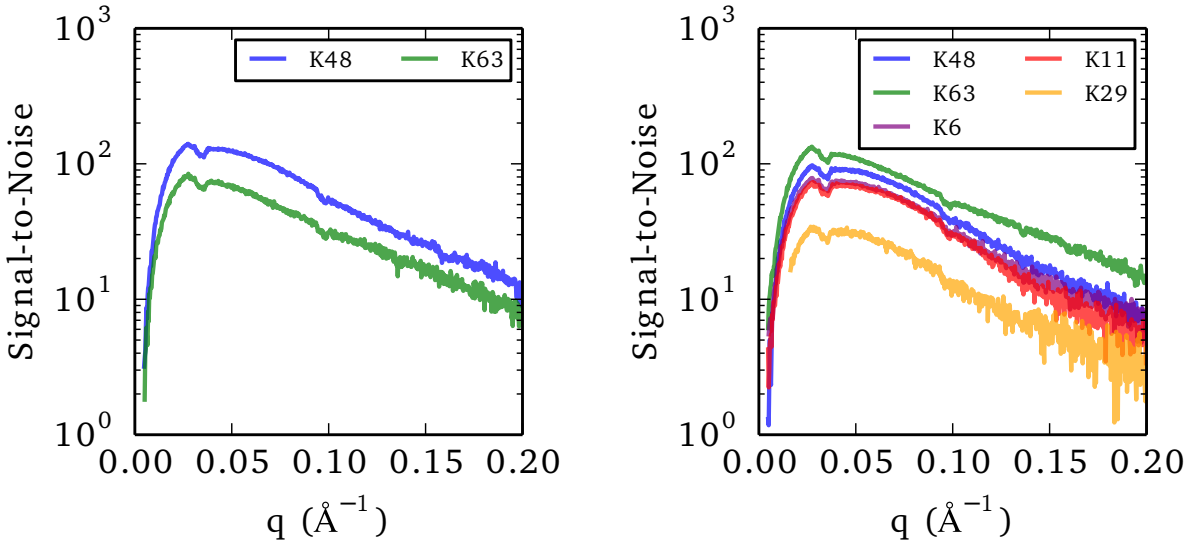

Figure S2: Signal-to-Noise ratios of the natively (left) and non-natively (right) linked tri-ubiquitin systems. The points in low q with low signal-to-noise were filtered out as the results of beam smearing and points above q = 0.2 Å^{-1} were removed due to theoretical limitations.

K6, non-native linkage

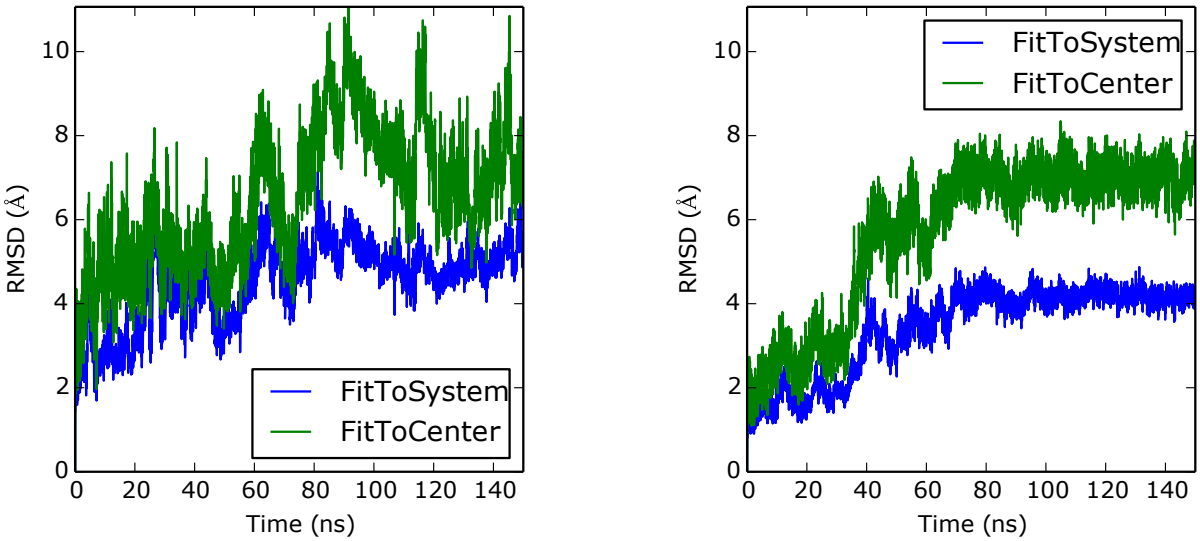

Figure S3: Backbone RMSD values for aMD (left) and cMD (right) simulations of the non-native K6 linkage. "FitToSystem" defines the backbone RMSD after least-squares fitting the whole backbone to the first frame, and the "FitToCenter" describes the backbone RMSD after least-squares fitting the central member backbone to the first frame. In this manner, "FitToSystem" describes total system mobility, while "FitToCenter" describes the flexibility of the distal groups. Thus, it becomes apparent that the distal groups in the aMD simulation (left, green) are significantly more mobile than in the cMD simulation (right, green).

K11, non-native linkage

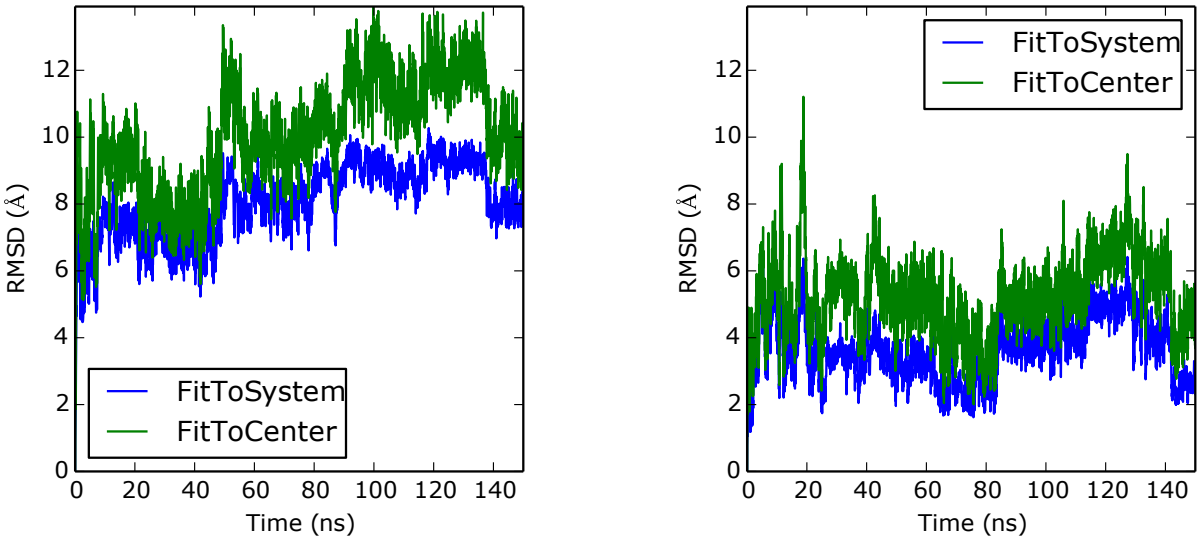

Figure S4: Backbone RMSD values for aMD (left) and cMD (right) simulations of the non-native K11 linkage.

K29, non-native linkage

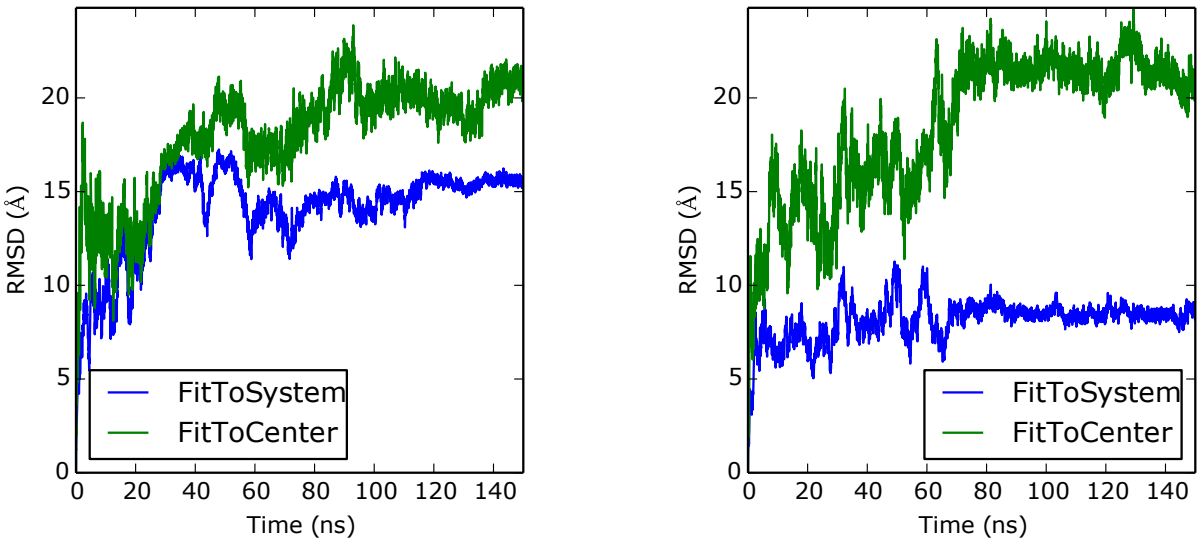

Figure S5: Backbone RMSD values for aMD (left) and cMD (right) simulations of the non-native K29 linkage.

K48, non-native linkage

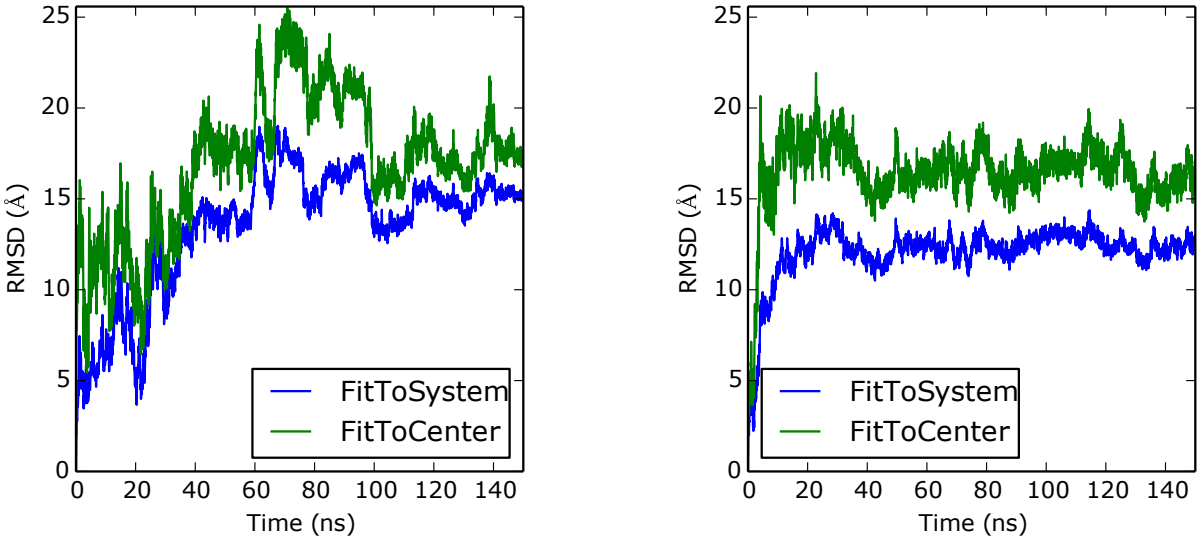

Figure S6: Backbone RMSD values for aMD (left) and cMD (right) simulations of the non-native K48 linkage.

K48, native linkage

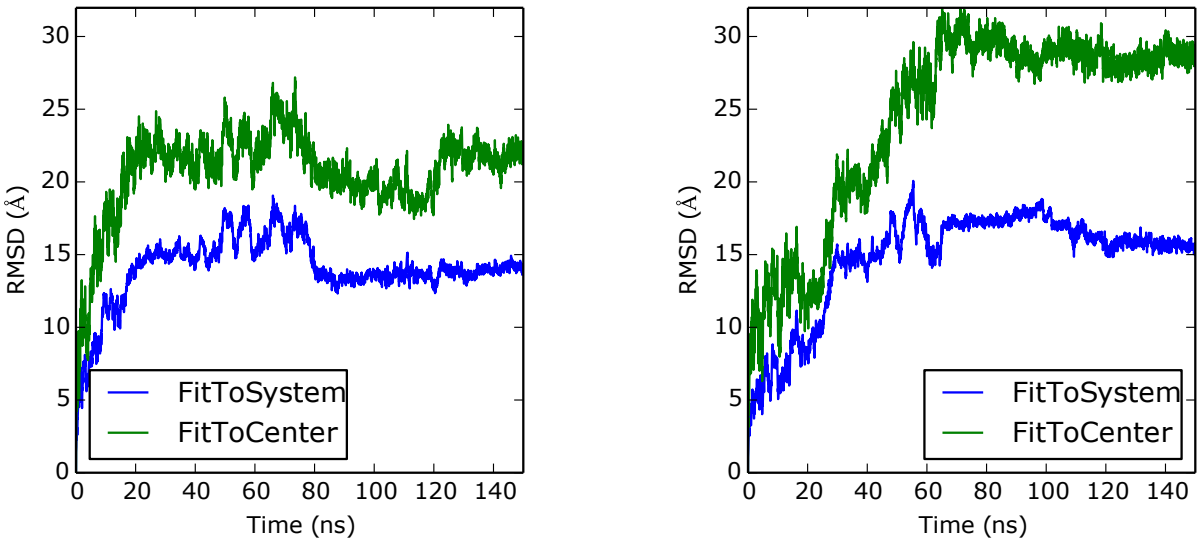

Figure S7: Backbone RMSD values for aMD (left) and cMD (right) simulations of the non-native nK48 linkage.

K63, non-native linkage

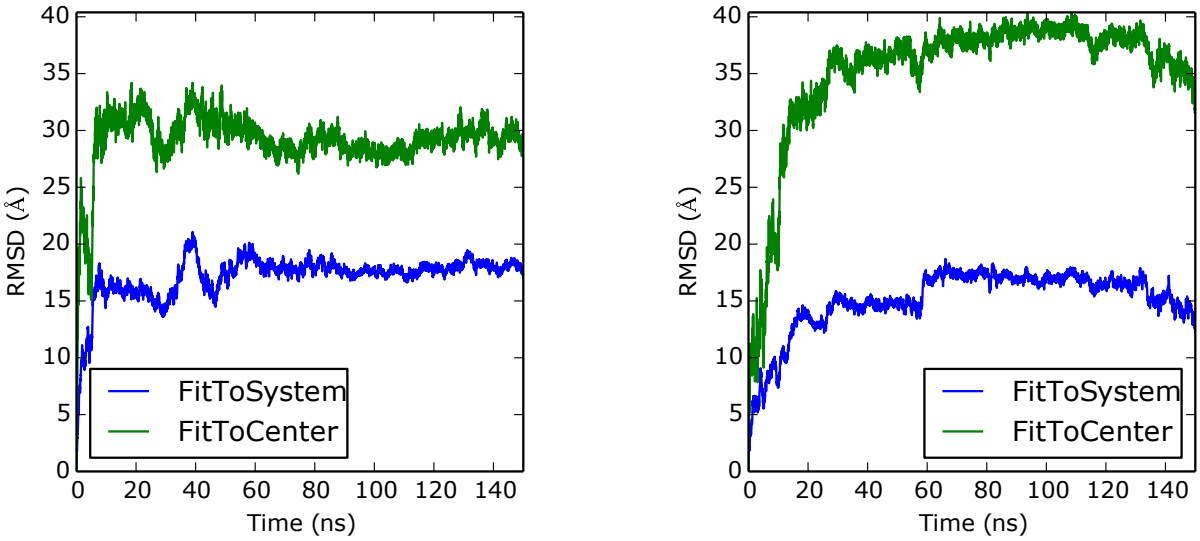

Figure S8: Backbone RMSD values for aMD (left) and cMD (right) simulations of the non-native K63 linkage.

K63, native linkage

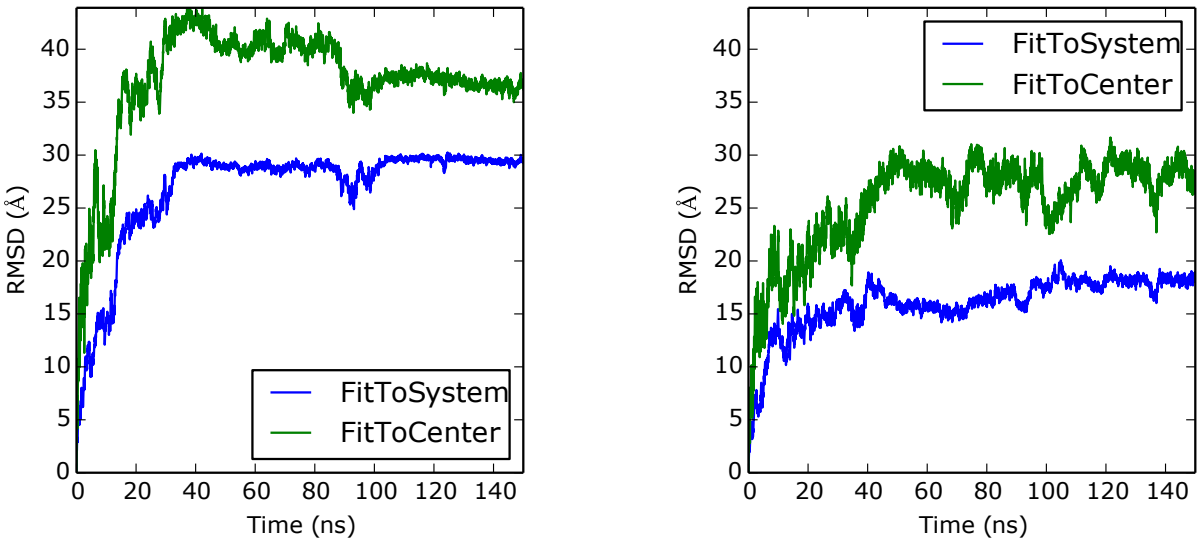

Figure S9: Backbone RMSD values for aMD (left) and cMD (right) simulations of the non-native nK63 linkage.

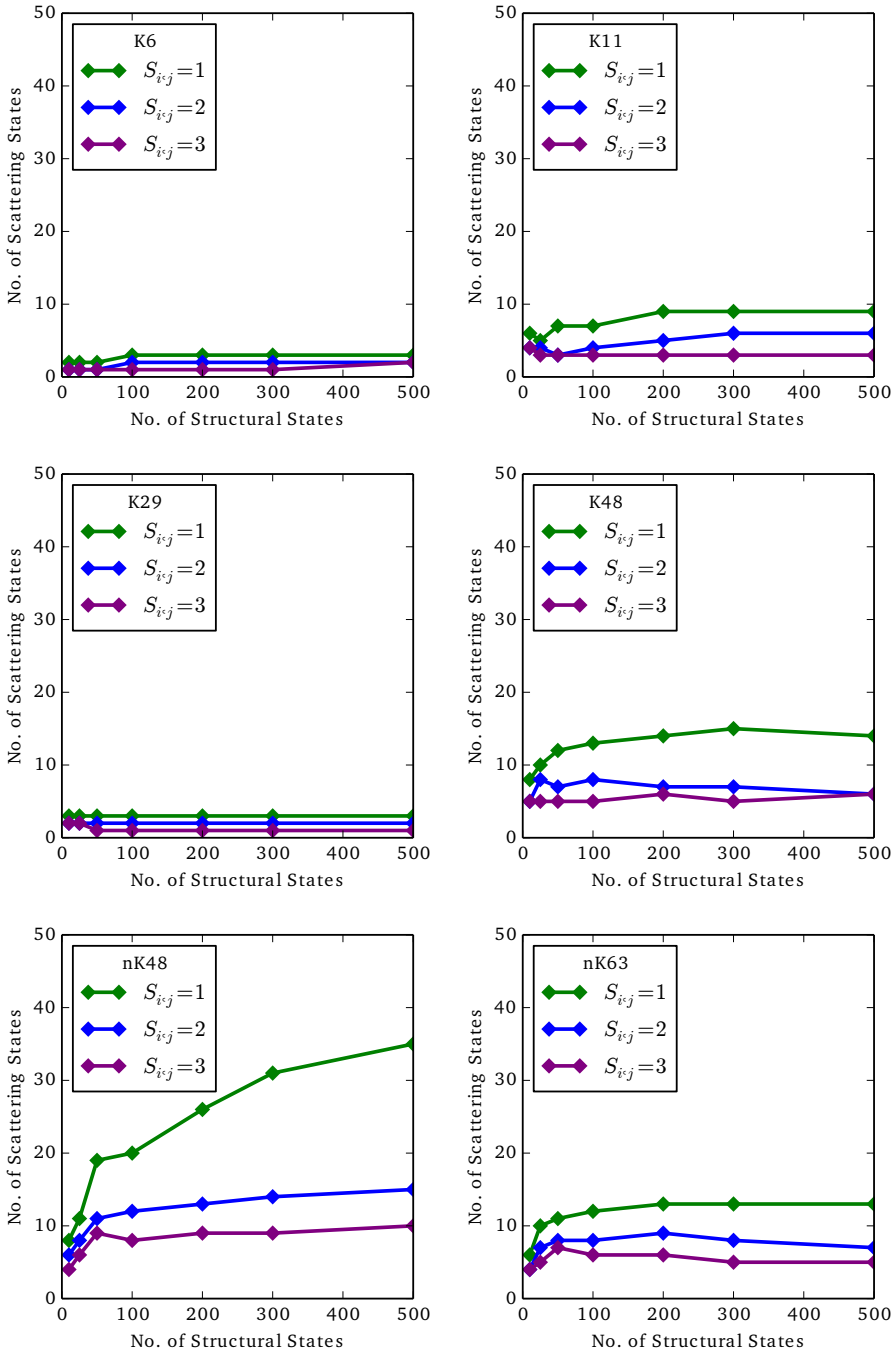

Figure S10: The number of scattering states vs initial structural clusters for the six systems not shown in the main text. nK63 and nK48 denote the systems with the native isopeptide linkage, and all other systems possess the non-native thiolene linkage between ubiquitin domains.

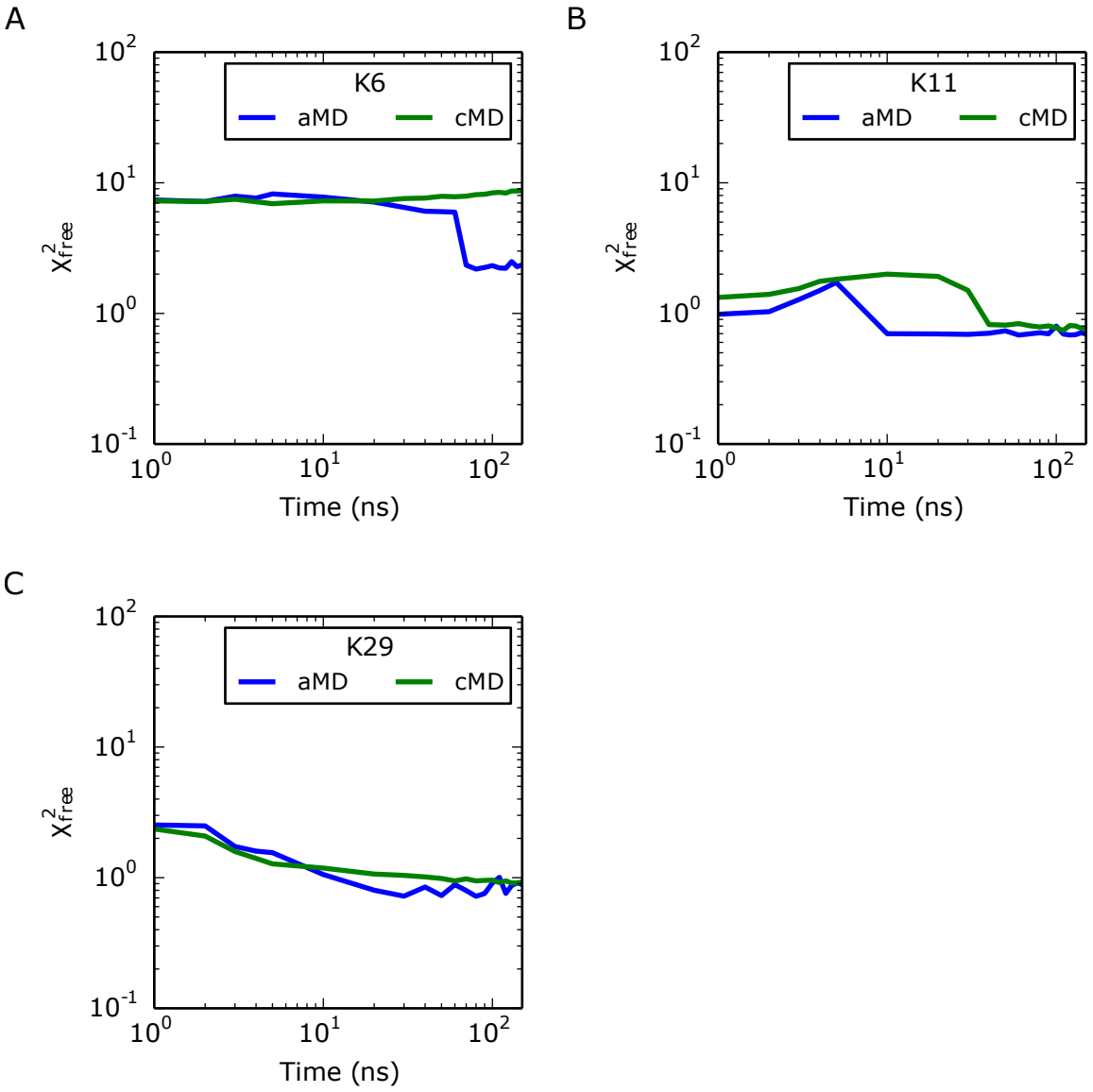

Figure S11: Reduced χ^2_{free} goodness-of-fit of identified ensembles vs sampling time for aMD (blue) and cMD (green) simulations of non-native (a) K6, (b) K11, and (c) K29. The overall quality of the initial conformations of K11 and K29 contributes to the comparable performance of the aMD and cMD simulations, but the K6 system displays a large disparity between the quality and convergence times of the aMD and cMD trajectories.

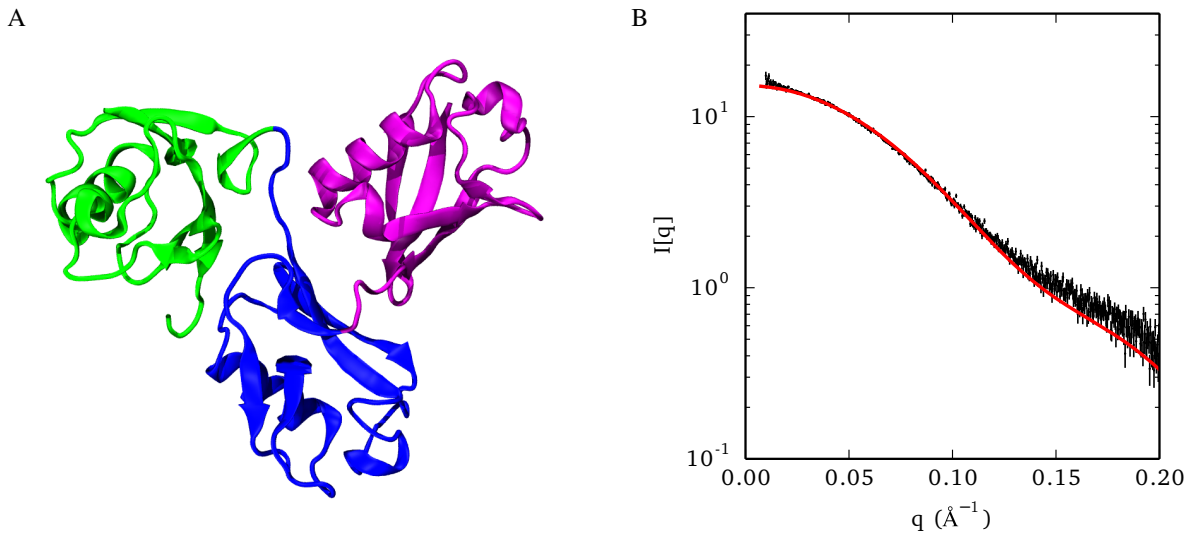

Figure S12: (a) Representative conformation of the population. (b) The ensemble scattering superimposed on the experimental curve.

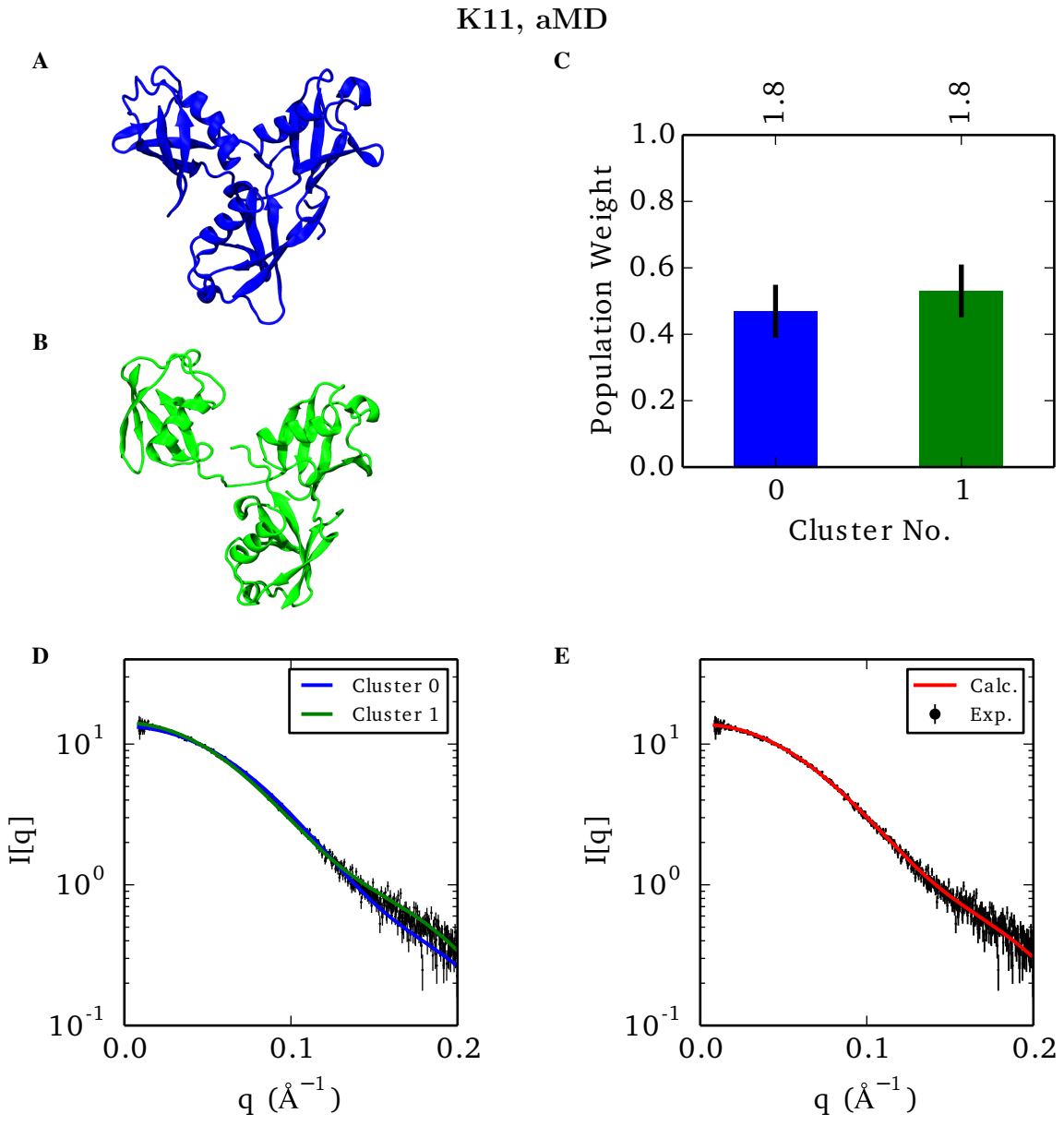

Figure S13: Representative conformations of the (a) compact and (b) slightly extended states. (c) The relative weights of each population and (d) their individual scattering curves. (e) The ensemble scattering superimposed on the experimental curve.

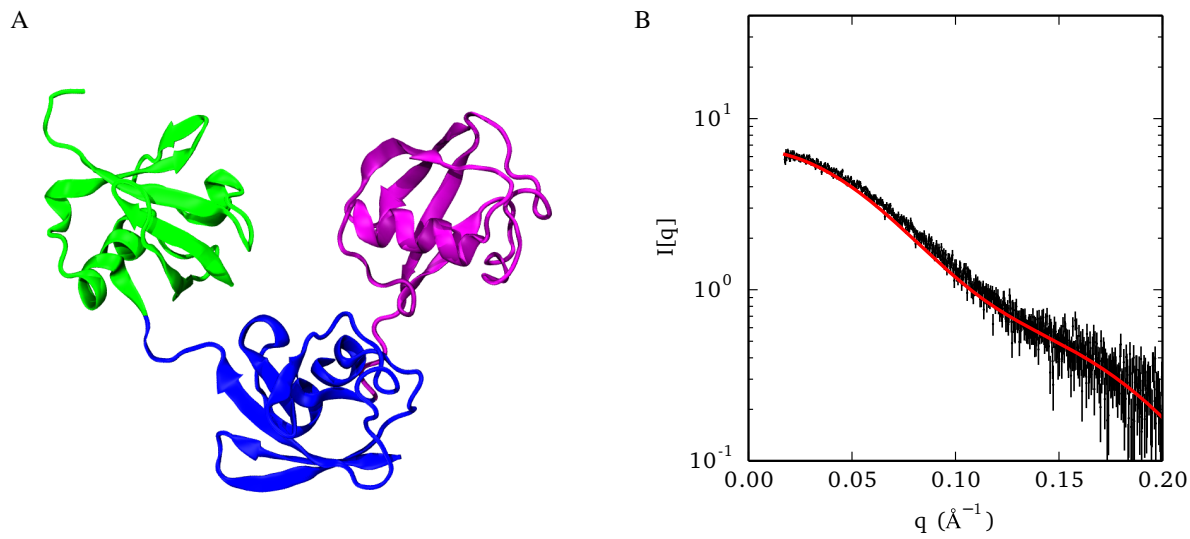

Figure S14: (a) Representative conformation of the population. (b) The ensemble scattering superimposed on the experimental curve.

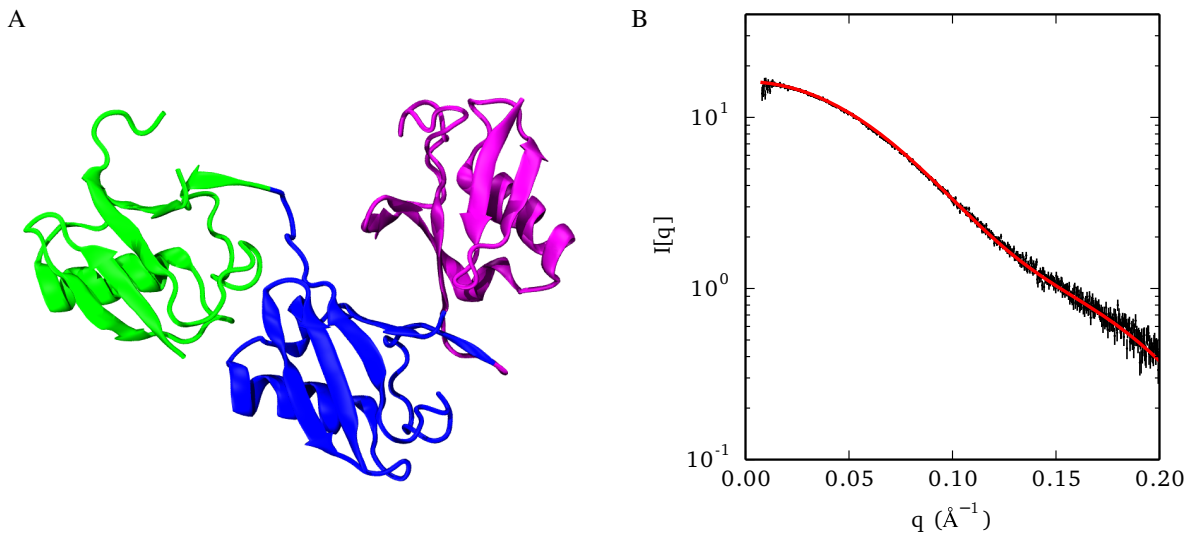

Figure S15: (a) Representative conformation of the population. (b) The ensemble scattering superimposed on the experimental curve.

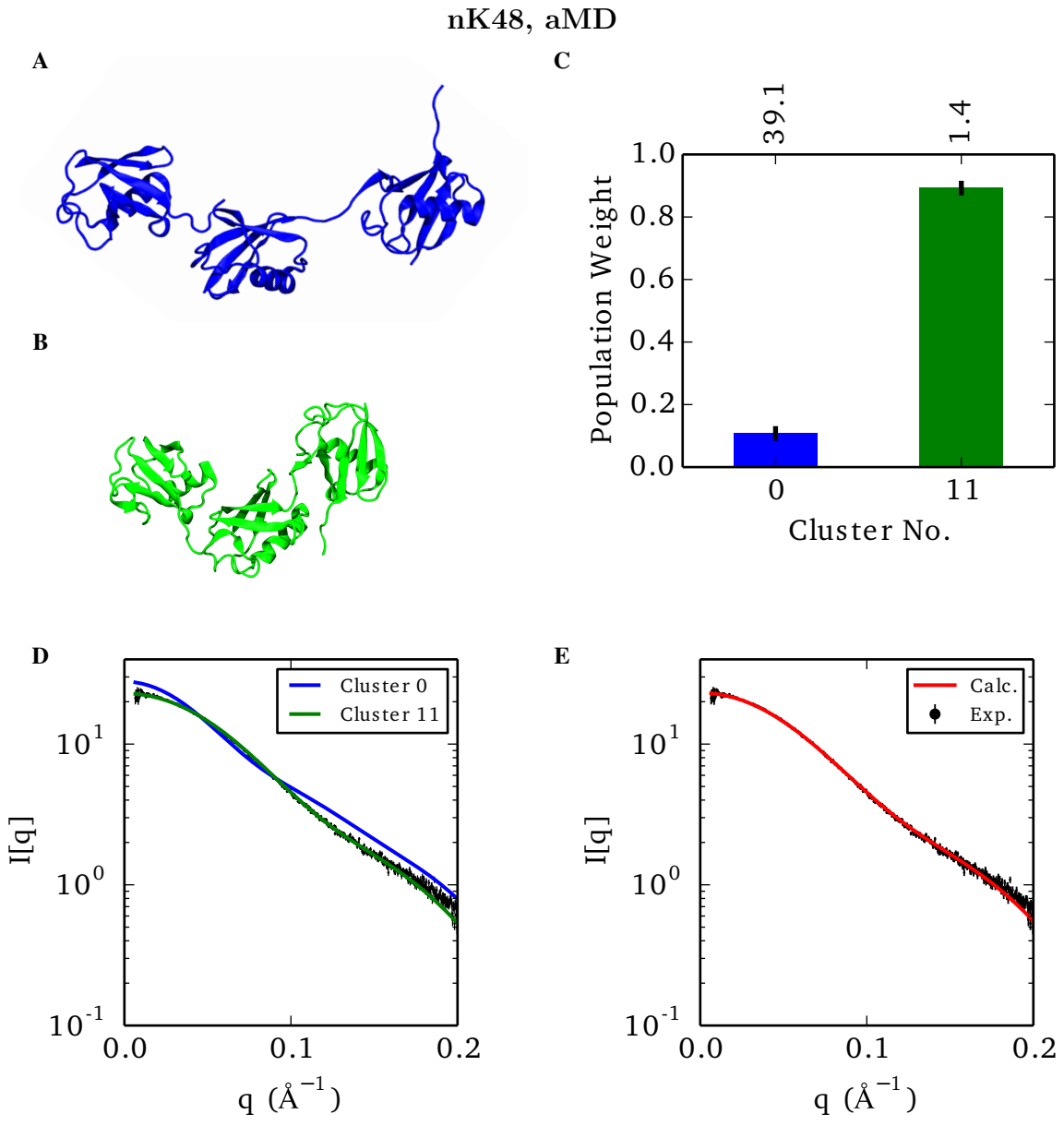

Figure S16: Representative conformations of the (a) extended and (b) compact states. (c) The relative weights of each population and (d) their individual scattering curves. (e) The ensemble scattering superimposed on the experimental curve.

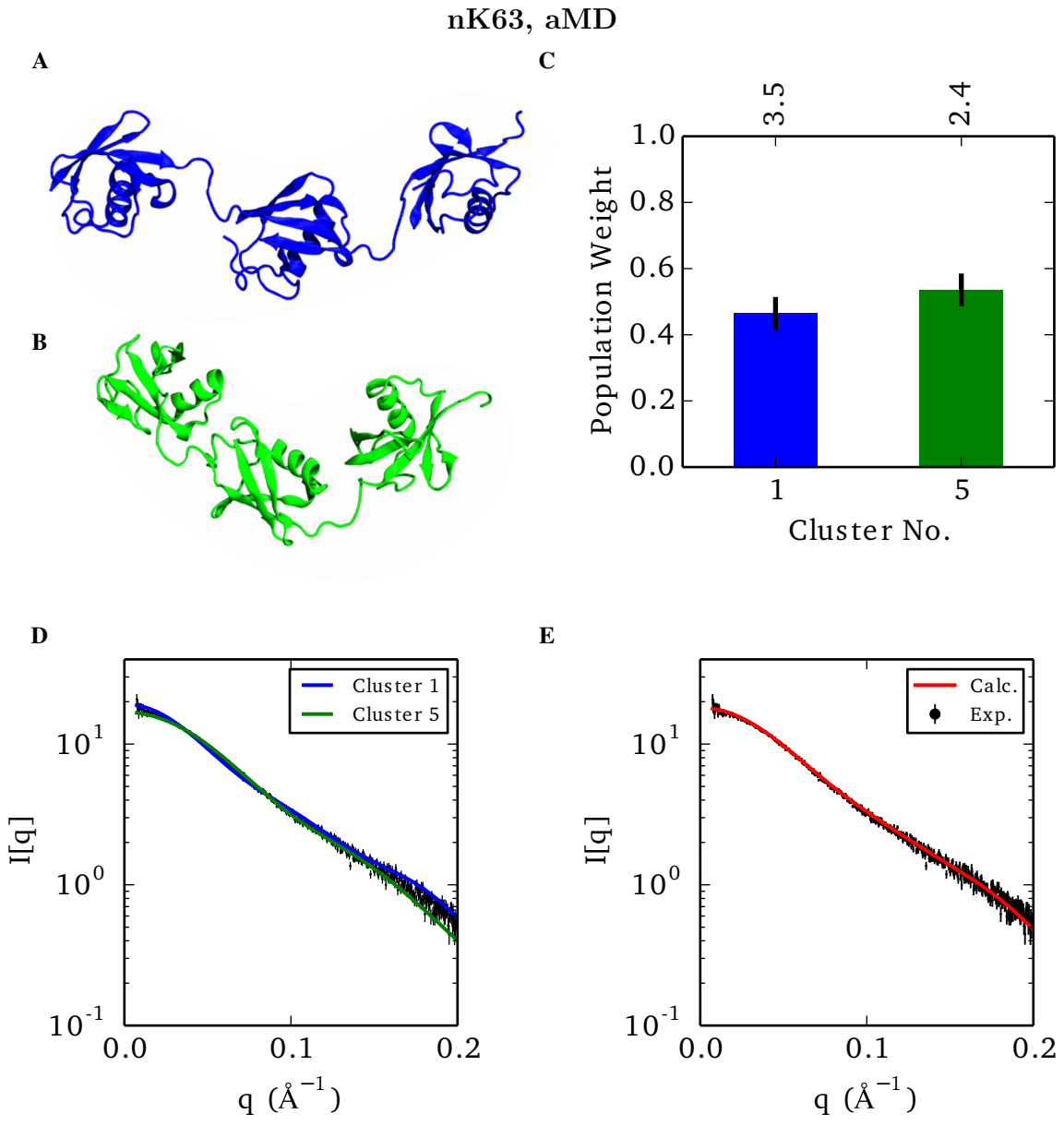

Figure S17: Representative conformations of the (a) extended and (b) compact states. (c) The relative weights of each population and (d) their individual scattering curves. (e) The ensemble scattering superimposed on the experimental curve.