Angewandte Chemie
Eine Zeitschrift der Gesellschaft Deutscher Chemiker

Supporting Information

Refinement of Highly Flexible Protein Structures using Simulation-Guided Spectroscopy

Jennifer M. Hays, Marissa K. Kieber, Jason Z. Li, Ji In Han, Linda Columbus, and Peter M. Kasson**

anie_201810462_sm_miscellaneous_information.pdf

SUPPORTING INFORMATION

Table of Contents

Experimental Procedures

mRMR-based selection of optimal DEER measurements

For each bacterial protein, we selected residue-residue pairs using the mRMR algorithm on 2μs MD ensemble simulations per protein. A Cα-Cα distance matrix was calculated using conformational snapshots at 500-ps intervals and these were histogrammed using 1-Å bins. Normalized histograms were used to calculate pair-configuration and pair-pair MI (Eqs. 1 and 2) as follows:

Typically, a conformation is represented as a 3N dimensional vector of atomic positions where N is the number of atoms. For selecting DEER pairs, however, a more natural choice of coordinate system is the set of distances between all possible residue-residue pairs. If the protein has *n* residues, there are (n^2-*n*)/2 possible pairs, and we can define the following conformation variable:

$$\vec{C} = (X_1 \quad X_2 \quad \cdots \quad X_i \quad \cdots \quad X_{(n^2-n)/2}),$$

where X_i is the distance between the i^{th} pair of residues. For mutual information calculations, these real-valued variables are then binned, such that each conformation variable is represented as a vector of integers, with each integer being a bin number. We thus have a set of observed conformations {c}.

In order to determine the most informative pairs, we calculate the mutual information (MI) between a pair X_i and the conformation variable C:

$$I_i(X_i, C) = \sum_{\{x_i\},\{c\}} P(x_i, c) \log \frac{P(x_i, c)}{P(x_i)P(c)},$$

where $P(x_i, c)$ is the joint probability function of pair *i* and conformation *c* and $P(x_i)$ and P(c) are the marginal probability distribution functions of pair *i* and conformation *c*, respectively.

An ordered list of highest-ranking mRMR pairs was then generated using greedy mRMR selection[1] (Table S1). Code implementing mRMR selection of residues for DEER experiments is available from: https://github.com/kassonlab/mRMR-DEER. The implementation also provides the ability to exclude user-defined residue-residue pairs, such as residues where spin label placement might disrupt function, but that feature was not needed here.

Setup and equilibration of MD simulations.

FhuA

Because the Ton box motif is highly mobile and thus poorly resolved with NMR and X-ray crystallography, no full-length apo structures of FhuA exist. We therefore used a previously published ensemble modeled using NIH-XPLOR to initialize our simulations[2]. This ensemble incorporated a set of MTSL spin-labels. The spin-labels were removed via a homology model with an incomplete apo structure (PDB ID 1BY3)[3]. The final full-length apo structure was inserted into a membrane of 756 DLPC lipids using the Gromacs tool g_membed.

In order to improve sampling of the heterogeneous Ton box motif, we ran an initial pulling simulation to extend the N-terminal domain into the periplasm. The simulation incorporated four harmonic, pairwise restraints between the C_β of the residue pairs 13-161, 13-228, 13-373, and 13-663. Each residue pair was pulled to a distance of approximately 5 nm over the course of 12 ns. This short simulation time is reasonable since this was intended only to generate initial states. Conformations were then sampled every ns to obtain 12 structures for subsequent unrestrained simulations. Finally, a brief 100-ps equilibration was run on each of the structures using the NPT conditions described in *Production MD Simulations* below. The final ensemble consisted of two replicates of these 12 states for a total of 24 ensemble members.

OprG

The 20 lowest energy structures previously identified (PDB ID 2N6L) were chosen as initial states. They were inserted into a DLPC membrane as follows: first, CHARMM-GUI was used to equilibrate a single OprG state obtained from the Orientations of Proteins in Membranes (OPM) database[4]. Then, each of the 20 low energy structures was aligned to the β-barrel of this single structure. Each system was solvated independently with approximately 40,000 TIP3P water molecules and ions were added to obtain a system with 150 mM NaCl and no net charge. The final systems were independently energy-minimized using steepest-descent for 5000 steps or until the largest force was less than 1000 kJ/mol/nm^2. Finally, a brief 100-ps equilibration was run using the NPT conditions described in *Production MD Simulations* below. Of these initial 20 systems, only six fully relaxed in the membrane; many of the initial loop

conformations extend downward into the plane of the membrane and thus are unlikely to be true conformational states[5]. The final ensemble consisted of four replicates of these six states for 24 total ensemble members.

Opa_{60}

The 20 lowest free-energy structures of Opa_{60} previously identified[6] (PDB ID 2MAF) were selected as initial states. Each Opa_{60} molecule was inserted into a membrane of 494 DMPC molecules as follows: the beta-barrel was aligned to previously embedded beta-barrel of a single structure from the Fox simulations. The protein and membrane were energy-minimized using the steepest-descent integrator for either 5000 steps or until the largest force was less than 1000 kJ/mol/nm^2, whichever occurred first. Each system was solvated independently with approximately 300,000 TIP3P water molecules, and ions were added to obtain a system with 150 mM NaCl and no net charge. The final systems were independently energy-minimized again using steepest-descent for 5000 steps or until the largest force was less than 1000 kJ/mol/nm^2. Finally, a brief 100-ps equilibration was run using the NPT conditions described in *Production MD Simulations* below.

Initial states for the second iteration of mRMR were obtained by resampling the mRMR-restrained ensemble simulations according to the joint distribution of the underlying DEER distributions (the individual distributions were assumed to be independent). The solvation, energy minimization, and initial equilibration protocols were identical to those of the ensembles described above.

Production MD Simulations

All production simulations were performed using a modified version of Gromacs 5.2 available at https://github.com/kassonlab/reMD-gromacs-5.2 and the CHARMM36 forcefield[7,8]. Simulations were run under NPT conditions using the velocity-rescaling thermostat at 310 K with a 2-ps time constant and pressure maintained at 1 bar using the Parrinello-Rahman barostat with a 10-ps time constant[9]. Covalent bonds were constrained using LINCS, and long-range electrostatics were treated using Particle Mesh Ewald[10]. For each protein, ensemble simulations were run until a total of 2μs of data were collected.

Expression, purification, labeling, and refolding of Opa_{60}.

The opa60 gene was sub-cloned into a pET28b vector (EMD chemicals, Gibbstown, NJ) containing N and C terminal His_6 – tags. Cysteine residues were introduced at regions of interest on Opa using PIPE Mutagenesis, and gene sequencing confirmed the mutations (Genewiz Inc., South Plainfield, NJ). The pET28b vectors containing a mutated opa60 gene were transformed into BL21(DE3) E. coli cells, and cultures were grown in Luria-Burtani (LB) media. Opa protein expression to inclusion bodies was induced with 1 mM isopropyl-β-thio-D-galactoside (IPTG). Cells were harvested and resuspended in lysis buffer [50 mM Tris-HCl, pH 8.0, 150 mM NaCl, and 1 mM TCEP-HCl (tris(2-carboxyethyl)phosphine hydrochloride)]. Following cell lysis, insoluble fractions were pelleted and resuspended overnight with lysis buffer containing 8 M urea. Cell debris was removed via centrifugation and unfolded Opa proteins in the soluble fraction were purified using Co^{2+} immobilized metal affinity chromatography, eluting in 20 mM sodium phosphate, pH 7.0, 150 mM NaCl, 680 mM imidazole, 8 M urea, and 1 mM TCEP. Purified Opa proteins were loaded on a PD-10 column (GE Healthcare Biosciences, Pittsburg, PA) to remove TCEP. Opa proteins were eluted with buffer (20 mM sodium phosphate, pH 7.0, 150 mM NaCl, and 8M urea) directly into five molar excess MTSL/R1 spin label [S-(2, 2, 5, 5-tetramethyl-2,5-dihydro-1H-pyrrol-3-yl)methyl methanesulfonothioate, Toronto Research Chemicals Inc., Toronto, Canada, stored as 100 mM stock in acetonitrile] for proteins containing a single cysteine and ten molar excess MTSL for Opa proteins with two cysteine residues. The proteins were spin labeled overnight at room temperature. Excess spin label was removed using a second PD-10 column, and the eluted protein was concentrated to approximately 150 - 200 μM. The labeled proteins were rapidly diluted 20-fold into 20 mM Tris-HCl, pH 8.0, 500 mM NaCl, 3 M urea, and 4.6 mM n-dodecylphosphocholine (FC-12, Anatrace), upon which Opa proteins fold into the detergent micelles over the course of three days at room temperature[6,11]. Folding efficiency was assessed with SDS-PAGE. Samples were dialyzed against 3 x 4L of 20 mM sodium phosphate, 150 mM NaCl for an hour each, removing any free spin. Opa proteins were concentrated to approximately 200 – 400 μM.

Double-electron electron resonance spectroscopy of Opa_{60} micelles.

Double-labeled Opa_{60} proteins in detergent micelles were measured using pulsed EPR with a Q-band Bruker E580 Spectrometer fitted with an ER5106-QT Q-band Flexline Resonator (Bruker Biospin) at 80 K. The spectrometer was connected to a 10W solid-state amplifier (Bruker AmpQ). All samples were prepared to a final protein concentration between approximately 100 and 200 μM with 10% deuterated glycerol. The samples were loaded into quartz capillaries with a 1.6 mm od x 1.1 mm id (Vitrocom) and flash frozen in liquid nitrogen. A four pulse DEER sequence was used with one 16 ns π/2, two 32 ns π observed pulses (at an observed frequency υ_1), and a π pump pulse (at a frequency υ_2) optimized at approximately 32 ns[12]. The pump frequency (υ_2) was set at the maximum of the nitroxide spectrum and the observed frequency (υ_1) is set to 75 MHz lower. Increasing inter-pulse delays at 16 ns increments were used with a 16-step phase cycle during data collection. Accumulation times were typically between 18 and 24 hours, with a dipolar evolution time between 2 and 3 μs. Dipolar evolution data were processed using DEERAnalysis2016 software[13] using Tikhonov regularization to generate distance distributions. Background subtraction of the distance distribution yielded error at each distance which was plotted as ranges representing fits that are within 15% root-mean-square-deviation of the best fit.

Restrained-ensemble biasing potentials.

To compare the quality of mRMR-guided versus spectroscopist-guided refinement of Opa_{60}, two ensemble refinements were run. The first incorporated experimental DEER distance distributions from high-ranking mRMR label pairs 31-166 and 88-162, while the second incorporated those from spectroscopist-selected label pairs 77-107 and 107-117. Restrained-ensemble biasing potentials

previously developed by Roux were applied to match MD distance histograms to DEER-derived distance distributions (Fig S3). Refinement was performed via restrained-ensemble simulation using a modified version of Gromacs 5.2 available at https://github.com/kassonlab/reMD-gromacs-5.2.

Both DEER-derived and MD-derived distance distributions were smoothed with a Gaussian filter. The smoothing parameter σ was chosen to reflect the experimental uncertainty in the fine modes of the DEER-derived distance distributions, 2 Å for the high-scoring mRMR pairs and 1 Å for the spectroscopist-selected pairs (SSP). Histograms were calculated using 1-Å bins.

Rather than updating the bias potential U_{bias} at every MD step, distance data were collected for all ensemble members for a period of 100 ps followed by a U_{bias} update. Additionally, a boxcar averaging filter was applied so that the simulation distance distributions were calculated using the last 10 ns of data for the first round of simulations and 25 ns for the second round of simulations. These modifications were implemented in order to obtain sufficient sampling for generating the MD distance distributions. Final distance distributions were calculated using the last 25 ns of data, while convergence monitoring using the Jensen-Shannon divergence was performed on a 10-ns window prior to the referenced time point (Fig. S3). An initial spring constant K=10 kJ/mol/nm2 was used for the first 40 ns in all three sets of simulations. After 40 ns, K was increased to 100 kJ/mol/nm2 in the mRMR-guided simulations in order to reverse the increase in J-S divergence observed from approximately 30-40ns.

Information-theoretic clustering

The final trajectories of both the mRMR-restrained and SSP-restrained ensembles were sampled at 0.5 ns intervals, and all Cα- Cα distances were calculated using Gromacs. Histograms of each Cα- Cα pair were constructed using 1-Å bins, and all pairwise mutual information values were calculated as:

$$I(X_1^{C\alpha}, X_2^{C\alpha}) = \sum_{\{x_1^{C\alpha}\},\{x_2^{C\alpha}\}} P(x_1^{C\alpha}, x_2^{C\alpha}) \ln \frac{P(x_1^{C\alpha}, x_2^{C\alpha})}{P(x_1^{C\alpha})\, P(x_2^{C\alpha})}.$$

Because closely related sets of pairs (high $I(X_1^{C\alpha}, X_2^{C\alpha})$) contain redundant information, it is possible to obtain an approximation of the Opa_{60} ensemble by knowing the distributions of only a subset of all Cα- Cα distances; that is, by grouping together sets of highly related pairs, one can obtain an approximation of the dimensionality ensemble. The quality of the approximation depends on how much information is lost by grouping together more and more diverse Cα- Cα pairs.

In order to quantitatively evaluate the dimensionality of the ensemble after incorporation of the mRMR or spectroscopist-selected pairs, we clustered closely related sets of Cα- Cα pairs using complete-linkage hierarchical clustering with an MI-based distance metric $D(X_1^{C\alpha}, X_2^{C\alpha}) = 1 - I(X_1^{C\alpha}, X_2^{C\alpha})/H(X_1^{C\alpha}, X_2^{C\alpha})$D($X_{C\alpha1}$, $X_{C\alpha2}$), where $H(X_1^{C\alpha}, X_2^{C\alpha})$ is the joint entropy of the pairwise Cα- Cα distance distributions:

$$H(X_1^{C\alpha}, X_2^{C\alpha}) = \sum_{\{x_1^{C\alpha}\},\{x_2^{C\alpha}\}} P(x_1^{C\alpha}, x_2^{C\alpha}) \ln\ P(x_1^{C\alpha}, x_2^{C\alpha}).$$

The maximum cluster diameter after each clustering step may be thought of as a measure of resolution, or quality of the approximation: as the cluster diameter increases, information about the ensemble is lost as increasingly more independent Cα- Cα pairs are grouped together and considered redundant.

The information-theoretic resolution is reported in Fig. 4 as 1-ε, i.e., 1 – max(cluster diameter).

Analysis of loop conformations:

Contact matrices were calculated for all inter-loop contacts in snapshots taken at 500-ps intervals using a distance cutoff of 6 Å. Principal components analysis was performed to obtain a new orthogonal basis set for loop-loop contacts. For restrained-ensemble simulations performed using mRMR-guided DEER data, all snapshots formed four compact and well-separated clusters in the subspace formed by the first three principal components. Similarly, for restrained ensemble simulations performed using SSP DEER data, all snapshots formed five well-separated clusters. These clusters and their corresponding centroids thus reflect the major contact modes between loops. This contact-matrix-based analysis was chosen because the loops are highly flexible, making the rigid-body alignment that underlies RMSD-based clustering less accurate.

Results and Discussion

Rank	mRMR Pairs	mRMR score	MI Pairs	MI score
1	36 171	4.110195	36 171	4.110195
2	91 165	3.203021	37 164	4.109671
3	25 167	3.156535	36 161	4.098168
4	39 158	3.141428	35 171	4.095264
5	85 170	3.105476	36 164	4.094354
6	32 163	3.117820	37 171	4.091122
7	36 91	3.099664	34 169	4.089396
8	34 168	3.121006	37 168	4.089102
9	94 167	3.106743	36 172	4.087823
10	38 154	3.096239	37 165	4.087248
11	85 163	3.105664	37 162	4.087004
12	39 164	3.112969	38 164	4.086218
13	36 175	3.091163	37 166	4.085327
14	30 167	3.094117	36 165	4.084580
15	89 158	3.090069	36 170	4.084023
16	39 171	3.082419	35 166	4.081497
17	91 173	3.084754	37 169	4.080652
18	26 163	3.078114	36 168	4.079957
19	35 95	3.074940	37 161	4.078506
20	35 165	3.089347	36 162	4.078061

Table S1. Ranking of top residue-residue pairs via mRMR and mutual information alone for Opa_{60}.

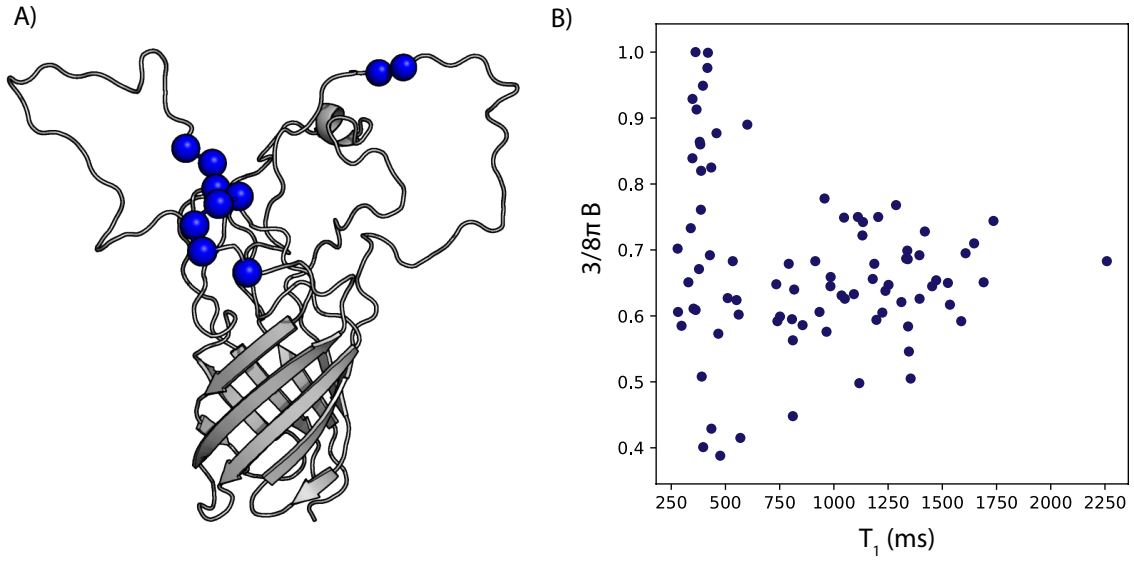


Figure S1. ENM-based scoring of flexibility correlates poorly with NMR data and identifies less informative loop regions. Elastic network models provide a computationally efficient means of approximating some protein motions. To assess this approach for Opa loop prediction and DEER pair selection, a Gaussian Network Model was used to predict Cα B-factors for Opa_{60}. The ten top-scoring residues are shown on the structure in (A). Many of the residues are located near the base of a single loop, while only two are located on a different loop in a more flexible region. Additionally, the ENM does not accurately reproduce the relative loop residue motion observed via NMR. The ENM-predicted B-factors correlate poorly with experimentally determined T_1 decays[2] (B); $r < 0.2$. The "high flexibility" residues identified by the ENM ends up closely resembling standard spectroscopist-guided pair-selection, with one residue in a region of high stability and one residue in a region of higher flexibility. Thus, mRMR-based pair selection on molecular dynamics trajectories, although computationally more expensive, yields more informative DEER pairs for Opa_{60}.

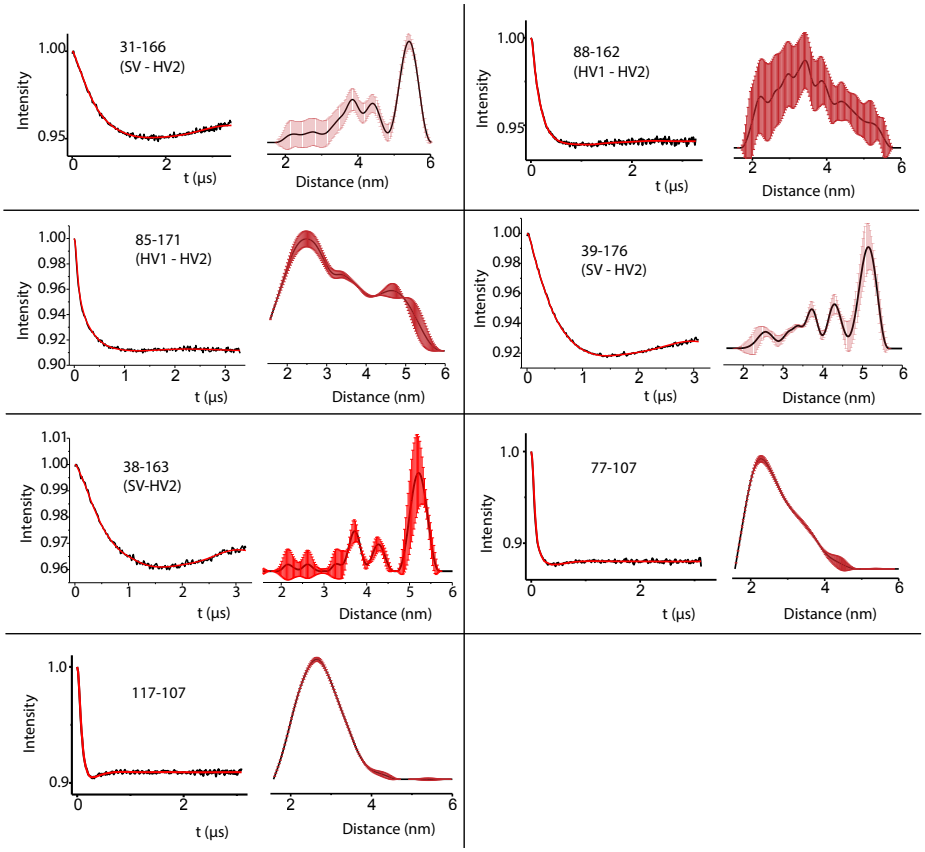
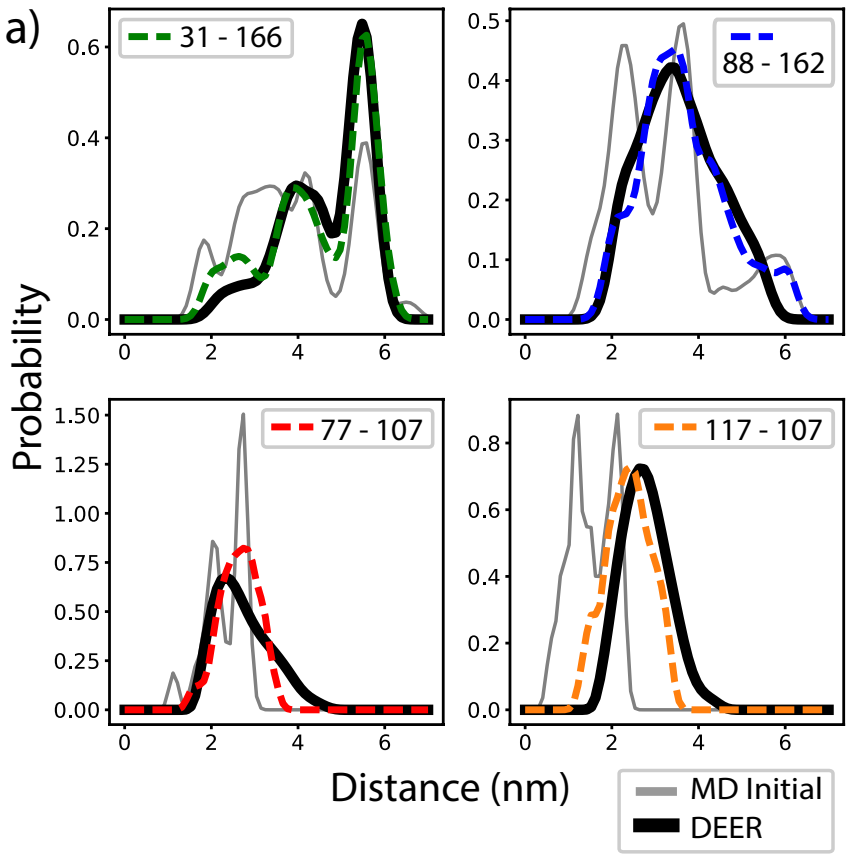


Figure S2. Measured spin-echo decays and fitted distributions. Fits are superimposed in red on the decays. The red error bars in the distance distributions represent uncertainty due to the background subtraction form factor that produce fits within 15% RMSD of the best fit.



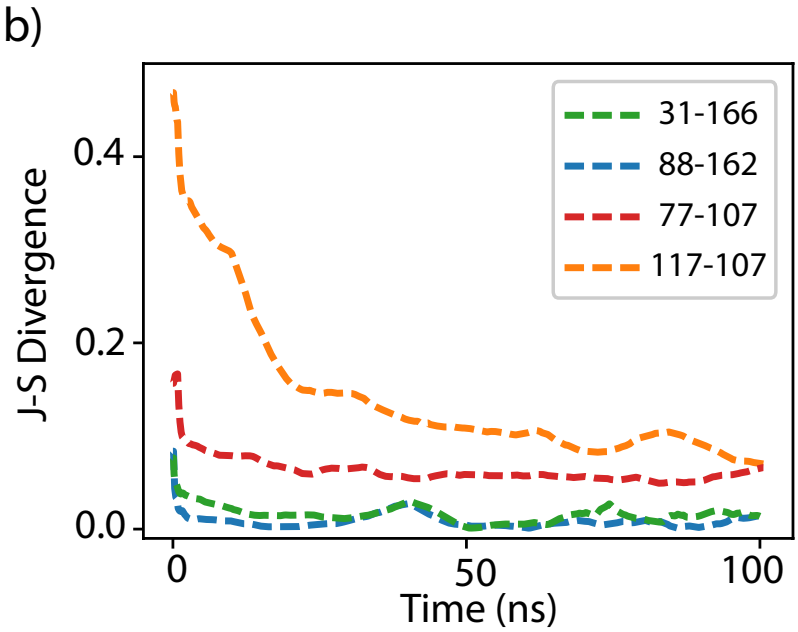
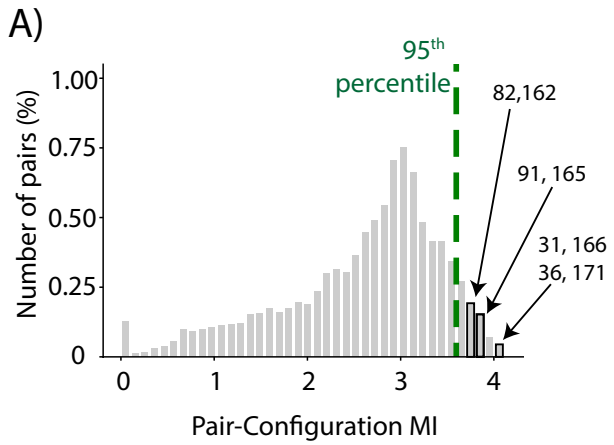


Figure S3: Restrained-ensemble simulations converge rapidly to experimental distributions. Convergence of restrained-ensemble simulations to DEER-derived distributions over 100 ns is plotted in (a) for both the high-scoring mRMR pairs and spectroscopist selected pairs. Convergence of both ensembles is quantified in (b) using Jensen-Shannon divergence.



B) mRMR scores:

Second Pair \ First Pair	36 - 171	31 - 166
91 - 165	3.203	3.137
88 - 162	3.108	3.171

Figure S4. Top mRMR-predicted pairs and measured pairs have near-identical pair-configuration mutual information and mRMR values. For operational reasons, the residue-residue pairs measured via DEER were slightly different than the top mRMR-predicted pairs. As shown in the histogram in a) and mRMR table b), the predicted and measured pairs are closely linked, having near-identical pair-configuration MI and mRMR scores. The mRMR table shows values of the mRMR statistic for the second residue-residue pair selected over all combinations of predicted and measured pairs. These statistics vary by less than 5%.

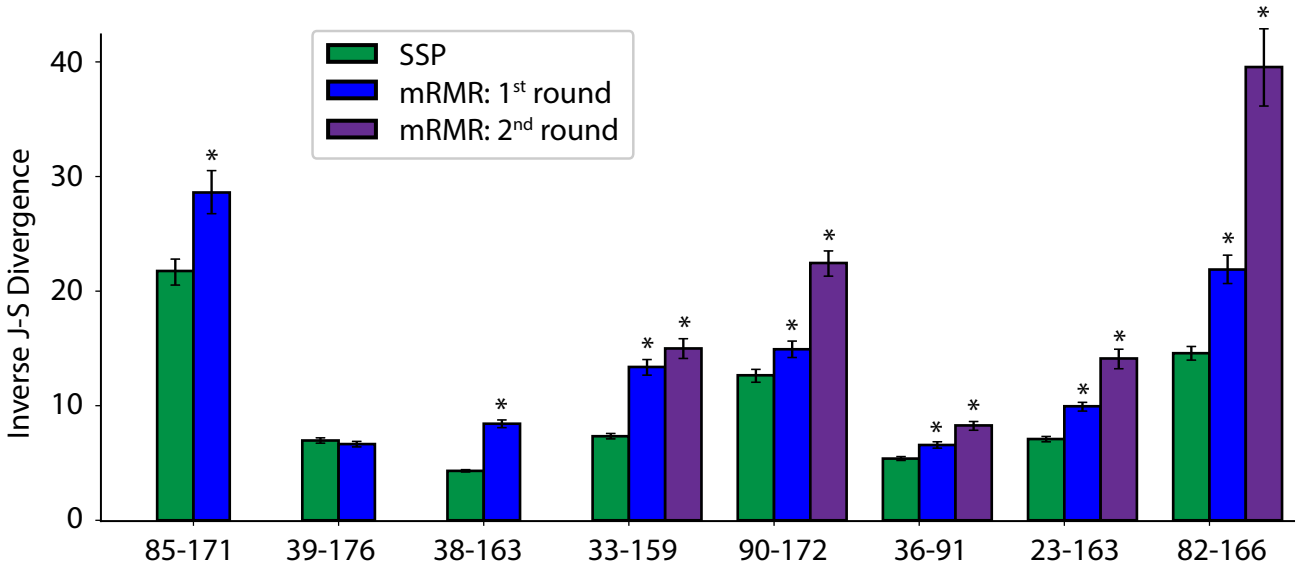


Figure S5: A second round of mRMR better refines the Opa_{60} conformational ensemble. Conformational ensembles refined using mRMR-selected pairs predict these new DEER distributions significantly better than conformational ensembles refined using spectroscopist-selected pairs (SSP) in seven of eight cases, quantified as inverse J-S divergences. Three of these DEER pairs were used for a second round of mRMR refinement; the resulting conformational ensemble outperforms both 1st-round ensembles in predicting the five pairs not used for refinement. Error bars represent 90% confidence using 1000 bootstrap replicates.

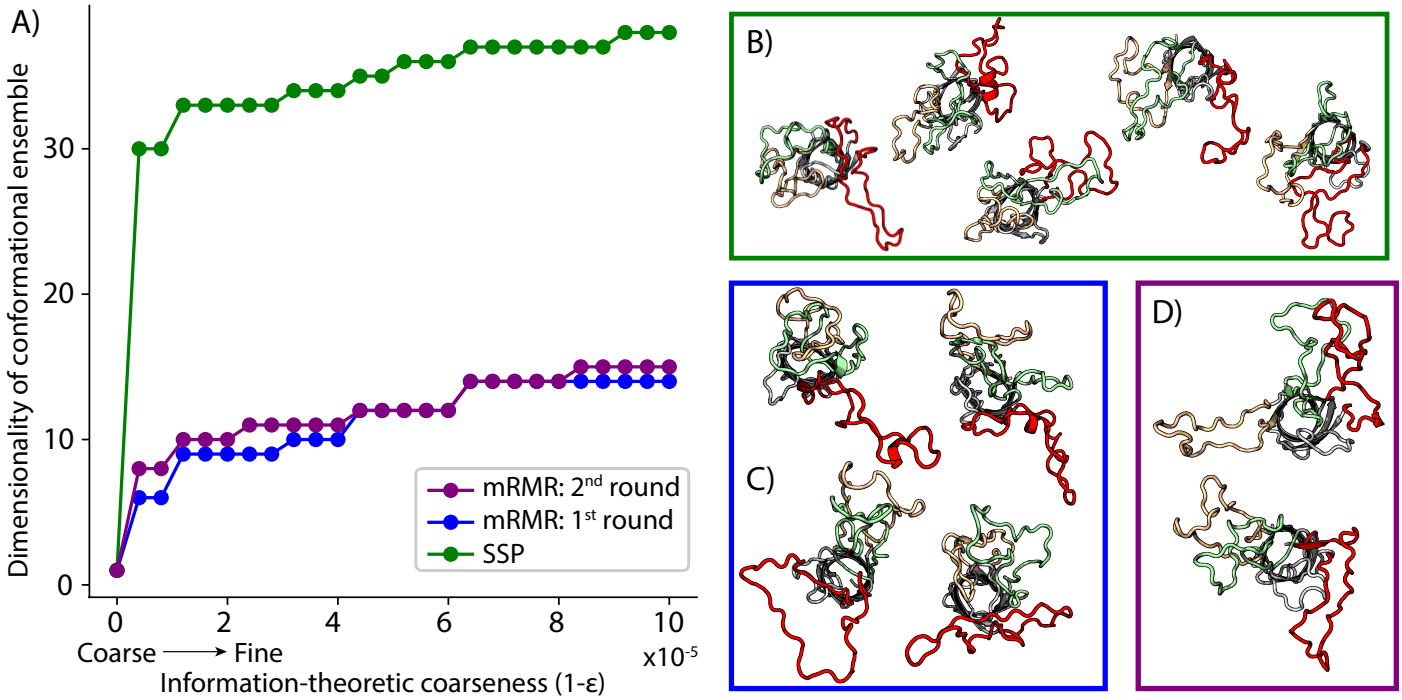


Figure S6: A second round of mRMR elucidates conformational heterogeneity of "two-and-one" loop configurations. The same "two-and-one" interaction patterns observed in the first round of mRMR-guided refinement (c) predominate in a second round of refinement (d). The conformational heterogeneity of the two-and-one interaction pattern is better resolved in the second round as evidenced by the additional single-loop extension in (d) and the unchanged dimensionality in (a). Conformational clusters from SSP-guided refinement are shown in (b) for completeness.

References

[1] H. Peng, F. Long, C. Ding, *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238.
[2] J. L. Sarver, M. Zhang, L. Liu, D. Nyenhuis, D. S. Cafiso, *Biochemistry* **2018**, *57*, 1045–1053.
[3] A. Šali, T. L. Blundell, *J. Mol. Biol.* **1993**, *234*, 779–815.
[4] M. A. Lomize, I. D. Pogozheva, H. Joo, H. I. Mosberg, A. L. Lomize, *Nucleic Acids Res.* **2012**, *40*, D370-6.
[5] J. Lee, D. S. Patel, I. Kucharska, L. K. Tamm, W. Im, *Biophys. J.* **2017**, *112*, 346–355.
[6] D. A. Fox, P. Larsson, R. H. Lo, B. M. Kroncke, P. M. Kasson, L. Columbus, *J. Am. Chem. Soc.* **2014**, *136*, 9938–9946.
[7] S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, et al., *Bioinforma. Oxf. Engl.* **2013**, *29*, 845–854.
[8] J. Huang, A. D. MacKerell, *J. Comput. Chem.* **2013**, *34*, 2135–2145.
[9] G. Bussi, D. Donadio, M. Parrinello, *J. Chem. Phys.* **2007**, *126*, 014101.
[10] T. Darden, D. York, L. Pedersen, *J. Chem. Phys.* **1993**, *98*, 10089–10092.
[11] D. A. Fox, L. Columbus, *Protein Sci.* **2013**, *22*, 1133–1140.
[12] M. Pannier, S. Veit, A. Godt, G. Jeschke, H. W. Spiess, *J. Magn. Reson.* **2000**, *142*, 331–340.
[13] G. Jeschke, V. Chechik, P. Ionita, A. Godt, H. Zimmermann, J. Banham, C. R. Timmel, D. Hilger, H. Jung, *Appl. Magn. Reson.* **2006**, *30*, 473–498.

Author Contributions

P.M.K designed and directed the research and co-wrote the manuscript. J.M.H developed the theory, performed simulations, analyzed data, and co-wrote the manuscript. M.K.K and L.C. designed and performed DEER experiments.