Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Elucidating the Role of Key Structural motifs in Antifreeze Glycoproteins

Poonam Pandey and Sairam S. Mallajosyula*

Department of Chemistry, Indian Institute of Technology Gandhinagar, Simkheda, Gandhinagar, Gujarat, India

E-mail: msairam@iitgn.ac.in

Phone: +91 - 79 - 32454998. Fax: +91 - 79 - 2397 2324

Supporting Information

Markov State Model (MSM) construction

Markov state model is a powerful and reliable approach to identify key conformations from MD simulation trajectories[1-3]. An MSM constitute of a network of conformation sets, extracted from simulation trajectories, with the transition probability matrix (T), characterizing the memoryless transition between states in a short time interval defined as lag time(τ). For a markov state model, time evolution of the system can be defined by the following equation

$P(t_0 + n\tau) = T^n(\tau)\, P(t_0)$

Where $P(t_0 + n\tau)$ and $P(t_0)$ are probability distribution of the states at time $t_0 + n\tau$ and t_0; and T(τ) is the transition probability matrix at the given lag time (τ).

The first step of MSM construction is featurization of raw MD simulation trajectories, using relevant matrices, i.e., cartesian coordinates of atoms, root mean square deviation, dihedral angles and distance between set of atoms. The next step involve the dimensionality reduction of input feature vectors using various approaches like principle component analysis (PCA) or time-structure independent component analysis[4-5] (tICA). Principle component (PCs) and time-structure independent component (tICs) are the linear combination of feature vectors having weight depicting the relevance of the corresponding vector. However, the difference in the above-mentioned projection approaches is that PCA is based on high variance linear combination of input feature vectors, whereas tICAs is based on high autocorrelation linear combination of the features. Once projected to tICA subspace, various clustering approaches, like k-means, k-centers can be used to obtain the discrete microstates, which will generate an assignment space, in which each conformation in the simulation trajectory will be assigned to the closest microstate. The next step involves the construction of MSM transition probability matrix P (τ) by counting the number of transitions among the microstates, using maximum likelihood approach.

To ensure the Markovian behavior of the model, multiple transition probability matrices can be constructed for different lag times and the relaxation timescales of the system can be estimated as

$$\tau_i = -\frac{\tau}{\ln \lambda_i(\tau)}$$

Where τ_i is the implied timescale corresponding to an aggregate transition between subsets of states in MSM, τ is the lag time, while $\lambda_i$is the eigen value of the transition probability matrix. If the model is Markovian, implied timescale must be independent of lag time[6].

In our study, we used backbone dihedrals matrices to generate initial feature vector, which were further subjected to tICA to project the input feature vectors to a low-dimensional subspace that best preserves the slowest conformational transitions. Once projected to the tICA subspace, distance-based clustering using the k-means algorithm, was performed to obtain the MSM metastable state definitions. The MSM transition matrix T (τ_{lag}) was estimated using the maximum likelihood method[7]. Implied timescales that plateau near the chosen lag time of 5 ns indicates that the chosen lag time is sufficient for MSM construction.

The microstates were further grouped together into Macrostates using robust Perron Cluster Cluster Analysis (PCCA) algorithm[8], grouping kinetically similar microstates. All models were built using MSMBuilder 3.3[7] and MDTraj 1.5[9] software package.

Table S1: Structural order parameters for the AFGP variants. Radius of gyration (R_g), end to end distance (R_{ee}), solvent accessible surface area (SASA) and dihedral RMSD.

AFGP Variant	$\langle R_g \rangle \pm \sigma$ (nm)	$\langle R_{ee} \rangle \pm \sigma$ (nm)	$\langle SASA \rangle \pm \sigma$ (nm^2)	$\langle$dihedral RMSD$\rangle \pm \sigma$ (°)
AFGP1	1.42 ± 0.11	2.84 ± 1.13	38.44 ± 1.80	22.39 ± 2.57
AFGP2	1.23 ± 0.13	2.72 ± 0.83	36.27 ± 1.40	28.73 ± 1.20
AFGP3	1.29 ± 0.17	2.93 ± 1.05	41.14 ± 1.85	29.58 ± 2.77
AFGP4	1.43 ± 0.11	3.46 ± 0.96	40.95 ± 1.34	29.27 ± 2.01
AFGP5	1.30 ± 0.15	2.43 ± 1.09	41.95 ± 2.53	31.57 ± 2.81

Table S2: The slow component of the relaxation time (τ_2) obtained from the Two-term exponential fitting parameters for the H-bond time autocorrelation functions of water-water H-bonds for the 1st and outer solvation shells around the five AFGP systems evaluated from 263 K, 268K, 271K, 273K and 298K simulations.

Temperature (K)		$\tau_2(ps)$					
		AFGP1	**AFGP2**	**AFGP3**	**AFGP4**	**AFGP5**	**Bulk Water**
263	**1st solvation shell (0 Å - 3.5 Å)**	24.14	15.37	18.29	13.42	17.51	7.63
	Outer solvation shell (10.0 Å – 12.0 Å)	8.27	8.62	8.55	8.38	9.01	
268	**1st solvation shell (0 Å - 3.5 Å)**	12.58	14.80	12.95	13.43	15.80	6.11
	Outer solvation shell (10.0 Å – 12.0 Å)	6.67	6.82	6.82	6.67	6.97	
271	**1st solvation shell (0 Å - 3.5 Å)**	17.67	13.62	11.78	10.61	13.49	5.79
	Outer solvation shell (10.0 Å – 12.0 Å)	6.07	5.93	6.01	5.84	5.97	
273	**1st solvation shell (0 Å - 3.5 Å)**	13.61	12.88	10.52	11.30	12.48	5.23
	Outer solvation shell (10.0 Å – 12.0 Å)	5.63	5.90	5.61	5.58	5.66	
298	**1st solvation shell (0 Å - 3.5 Å)**	6.70	5.78	5.44	4.88	14.26	3.03
	Outer solvation shell (10.0 Å – 12.0 Å)	3.08	3.12	3.21	2.83	6.00	

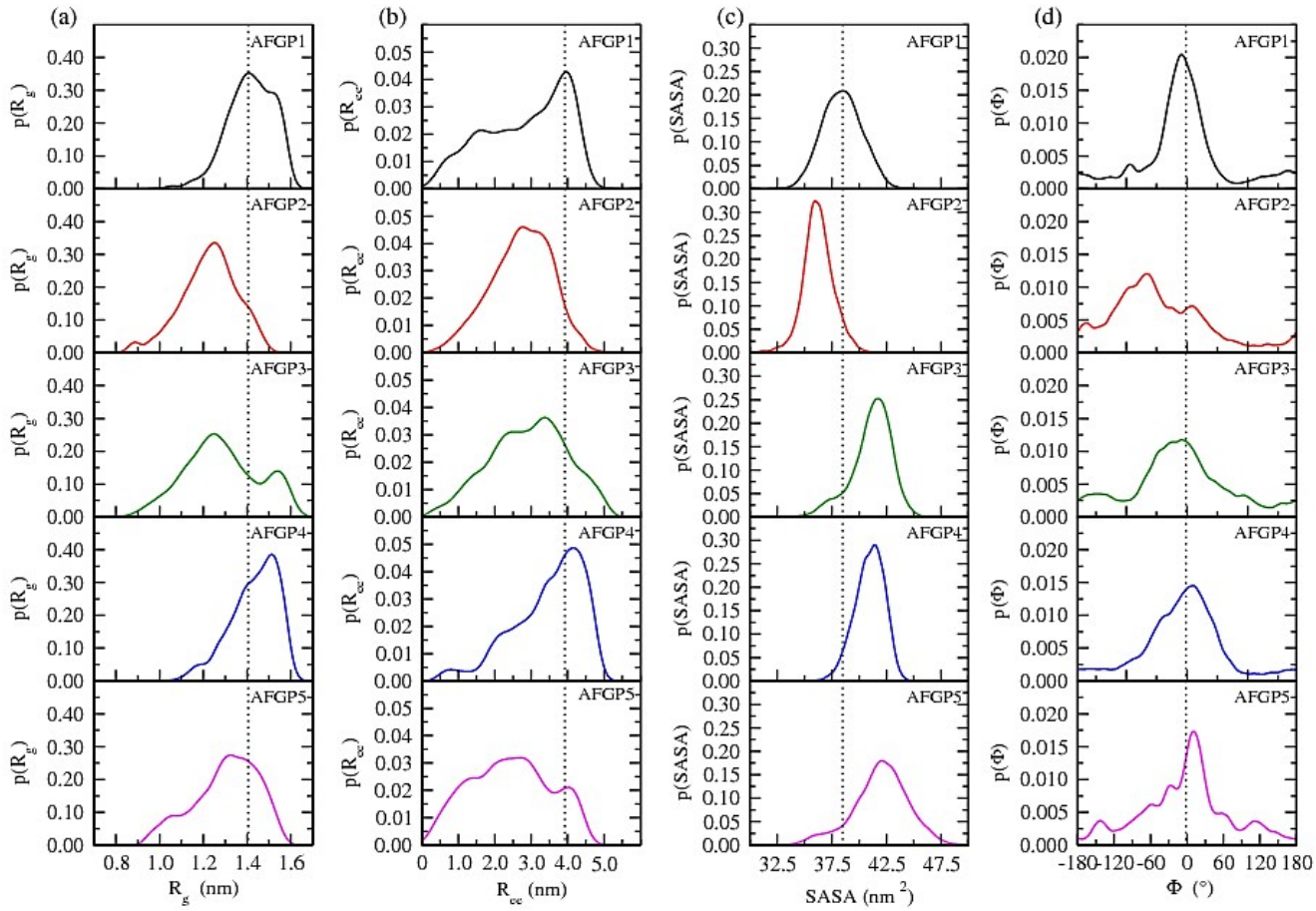


Figure S1. Density estimation of the (a) radius of gyration $p(R_g)$, (b) end to end distance $p(R_{ee})$ and (c) solvent accessible surface area p(SASA) and (d) pseudo dihedral angles for all the 5 AFGP variants, using a Gaussian kernel.

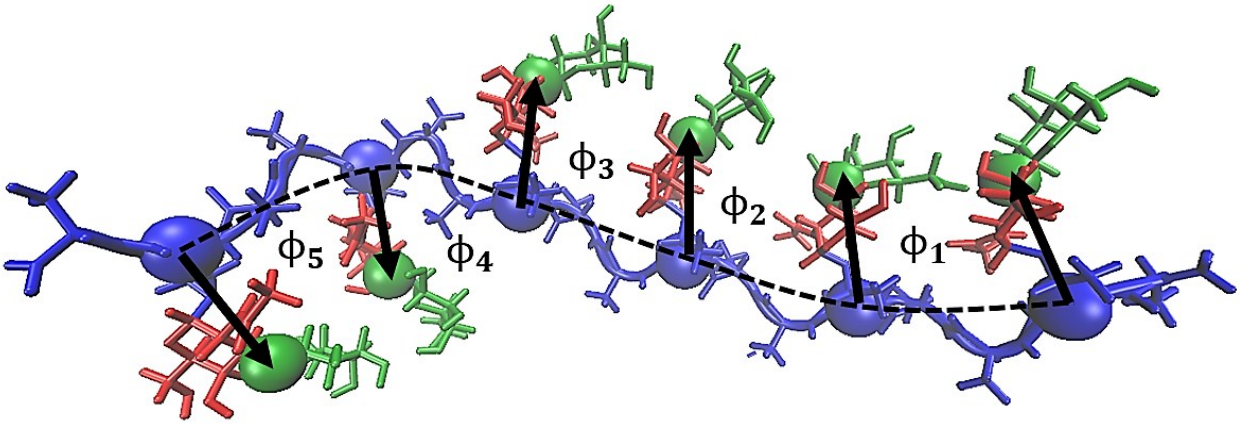
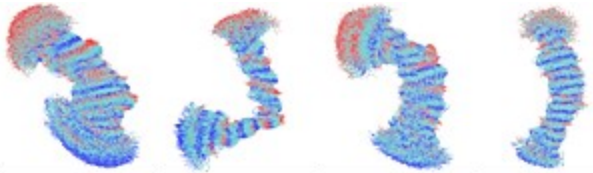


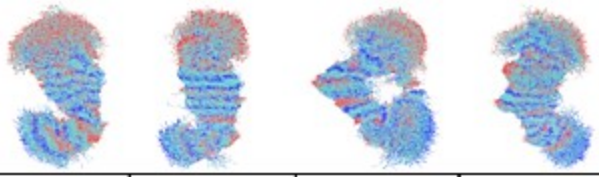
Figure S2. Graphical representation of pseudo dihedral angles ($\phi_1, \phi_2, \phi_3, \phi_4 \; and \; \phi_5$) defined as the dihedral angles formed between O3 oxygens of the first sugar attached to the protein, with the other two atoms being the Thr Cα atoms of the underlying peptide backbone, i.e. O3-Cα-Cα-O3.

(a) AFGP1



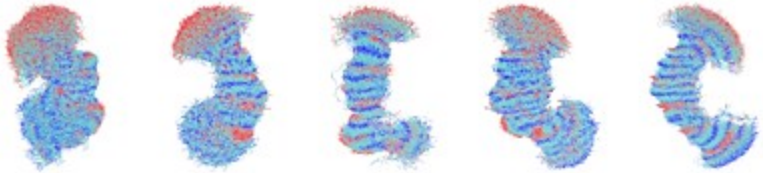
Fractional Population	0.70	0.20	0.05	0.05	-

(b) AFGP2



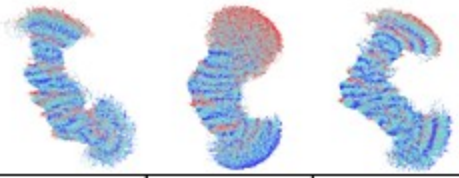
Fractional Population	0.29	0.23	0.26	0.22	-

(c) AFGP3



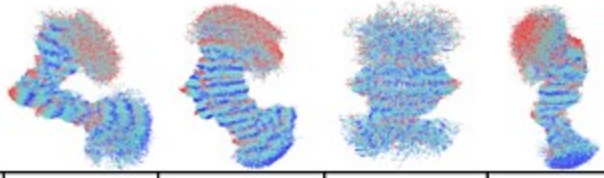
Fractional Population	0.33	0.23	0.06	0.25	0.13

(d) AFGP4



Fractional Population	0.07	0.86	0.07	-	-

(e) AFGP5



Fractional Population	0.18	0.52	0.05	0.25	-

Figure S3. Clustered conformation with Fractional population for all the five AFGP variants.

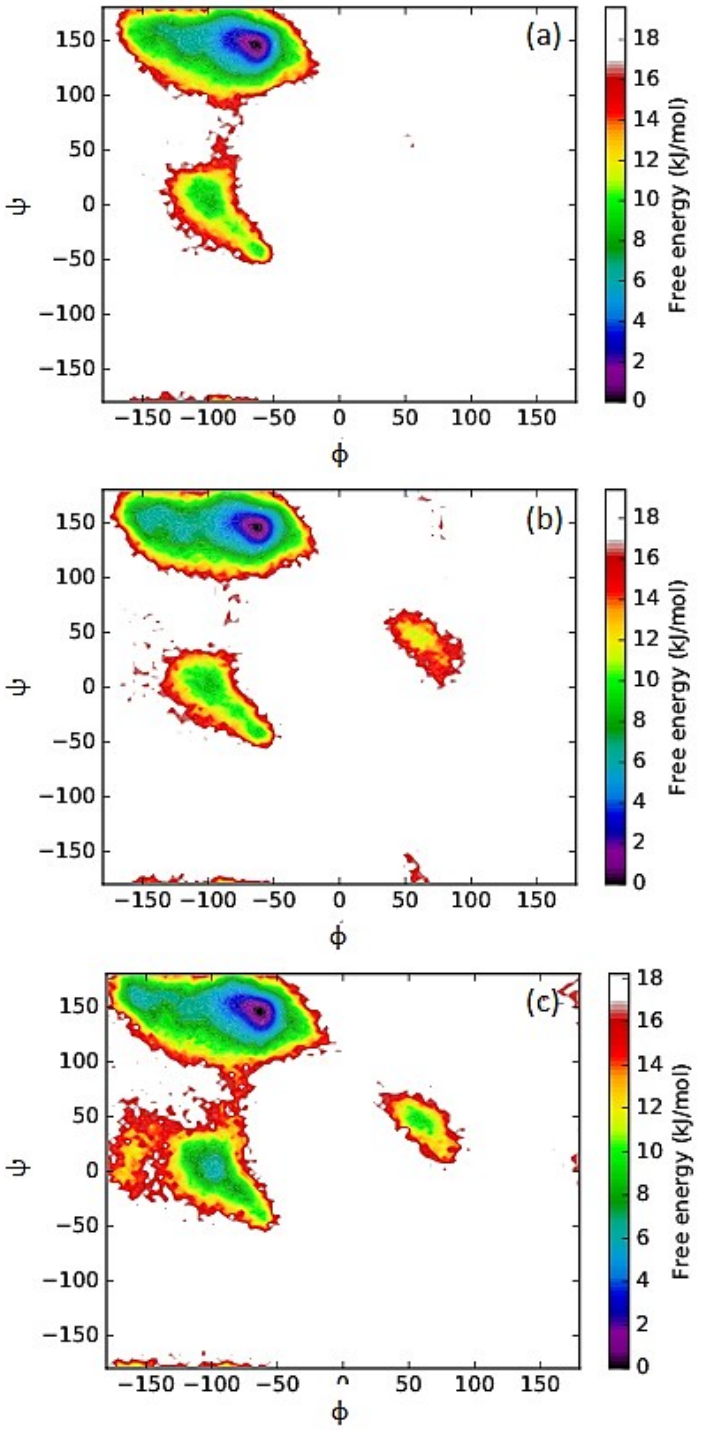


Figure S4. 2D relative free energy surfaces for the ϕ/ψ distributions corresponding to the peptide backbone obtained from the structures belonging to the most populated clusters for (a) AFGP1 (cluster1), (b) AFGP4 (cluster2) and (c) AFGP5 (cluster2).

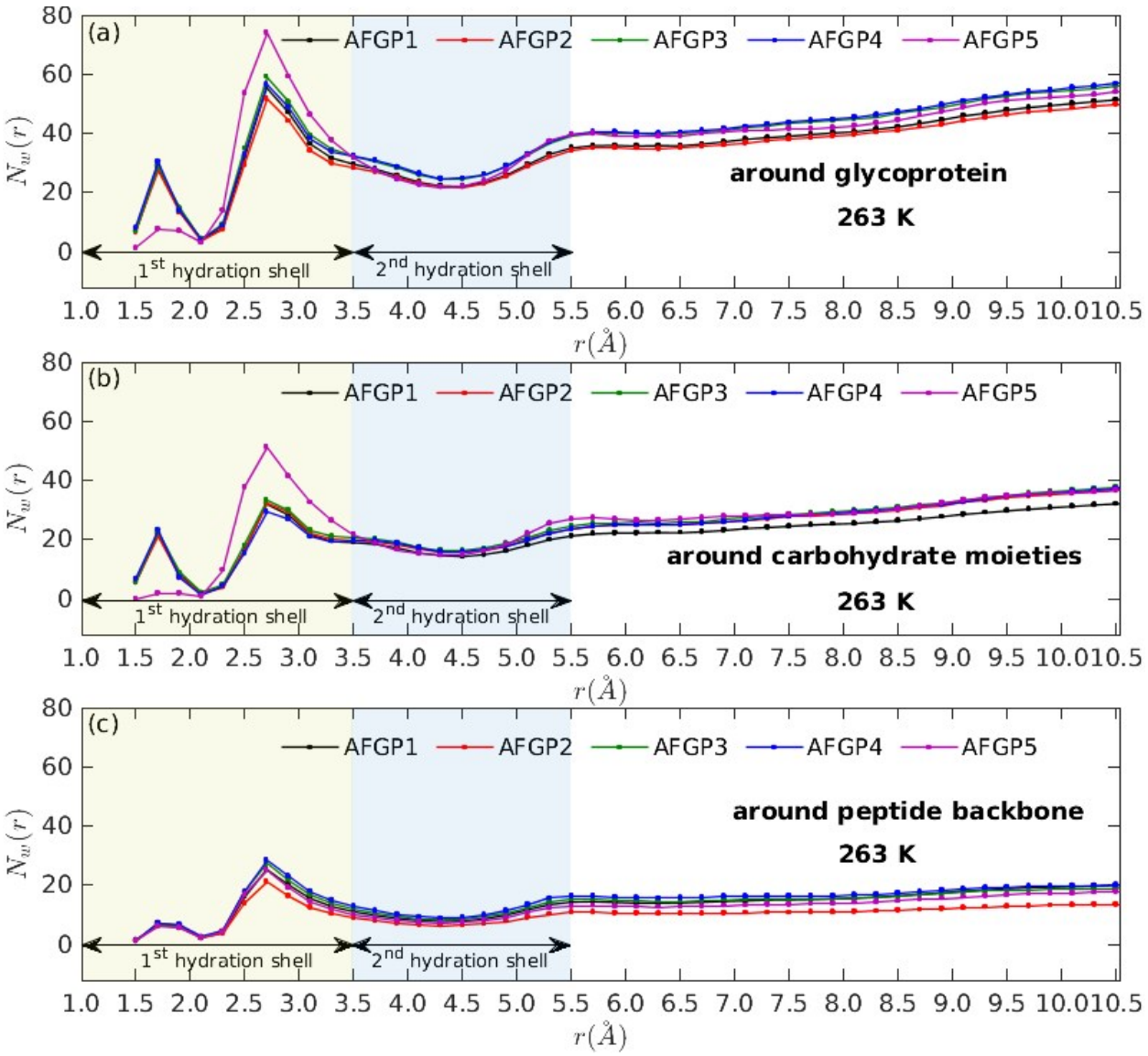


Figure S5. $N_w(r)$ distribution for water molecules as a function of distance evaluated in increments of 0.2 Å shells from (a) whole glycoprotein, (b) carbohydrates only and (c) around peptide backbone at 263 K.

References

1. Yao, Y.; Cui, R. Z.; Bowman, G. R.; Silva, D.-A.; Sun, J.; Huang, X., Hierarchical Nyström methods for constructing Markov state models for conformational dynamics. *The Journal of chemical physics* **2013,** *138* (17), 05B602_1.
2. Bowman, G. R.; Huang, X.; Pande, V. S., Using generalized ensemble simulations and Markov state models to identify conformational states. *Methods* **2009,** *49* (2), 197-201.
3. Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C., Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *The Journal of chemical physics* **2007,** *126* (15), 04B616.
4. Naritomi, Y.; Fuchigami, S., Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: the case of domain motions. *The Journal of chemical physics* **2011,** *134* (6), 02B617.
5. Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F., Identification of slow molecular order parameters for Markov model construction. *The Journal of chemical physics* **2013,** *139* (1), 07B604_1.
6. Swope, W. C.; Pitera, J. W.; Suits, F., Describing protein folding kinetics by molecular dynamics simulations. 1. Theory. *The Journal of Physical Chemistry B* **2004,** *108* (21), 6571-6581.
7. Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S., MSMBuilder2: modeling conformational dynamics on the picosecond to millisecond scale. *Journal of chemical theory and computation* **2011,** *7* (10), 3412-3419.
8. Deuflhard, P.; Weber, M., Robust Perron cluster analysis in conformation dynamics. *Linear algebra and its applications* **2005,** *398*, 161-184.
9. McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S., MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophysical journal* **2015,** *109* (8), 1528-1532.