Size-and-Shape Space Gaussian Mixture Models for Structural Clustering of Molecular Dynamics Trajectories.

Supporting Information.

Heidi Klem,† Glen M. Hocky,‡ and Martin McCullagh*,¶

†*Department of Chemistry, Colorado State University, Fort Collins, CO 80523*

‡*Department of Chemistry, New York University, New York, NY 10003*

¶*Department of Chemistry, Oklahoma State University, Stillwater, OK 74078*

E-mail: martin.mccullagh@okstate.edu

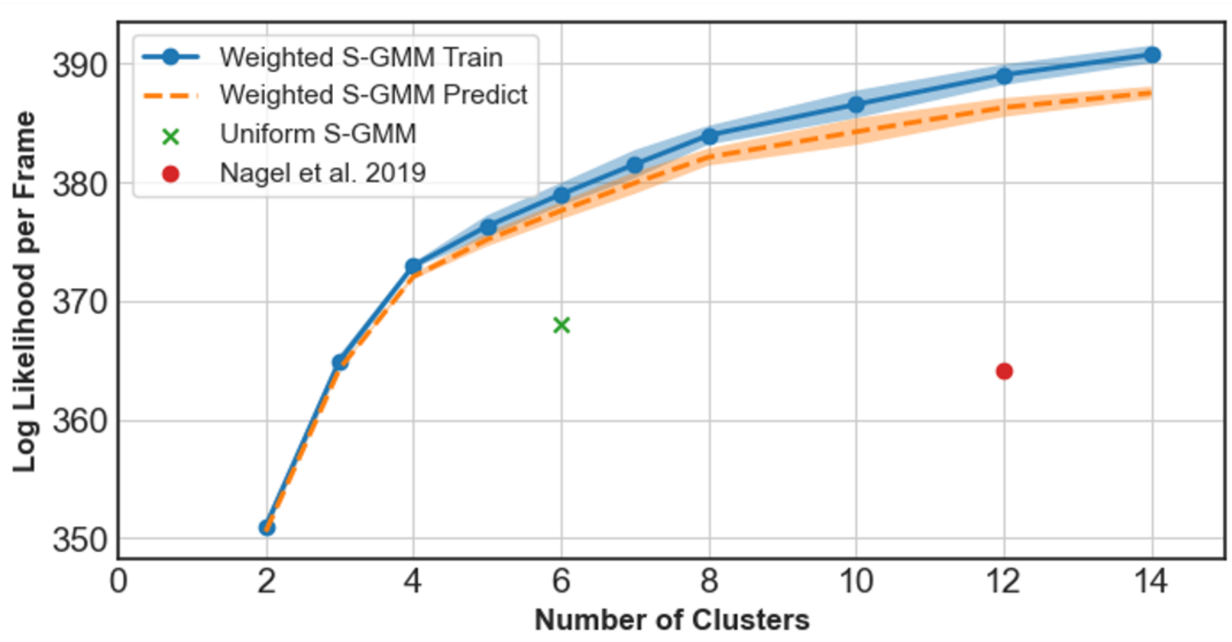

Figure S1: Log likelihood of weighted shape-GMM (S-GMM) as a function of number of clusters for HP35 trained on $\sim 25K$ frames (blue line) in a 10-fold cross validation (orange dashed) scheme. The shaded region indicates the 90% confidence interval. The features are C, CA and N backbone atoms of residues 2-34, the C atom of residue 1 and N atom of residue 35 (101 atoms total). The log likelihood value from a uniform 6-state shape-GMM is shown as a green x marker. The red circle marker indicates the log likelihood that results from the HP35 discretization of Nagel *et al*. clustering procedure resulting in 12-states.[2]

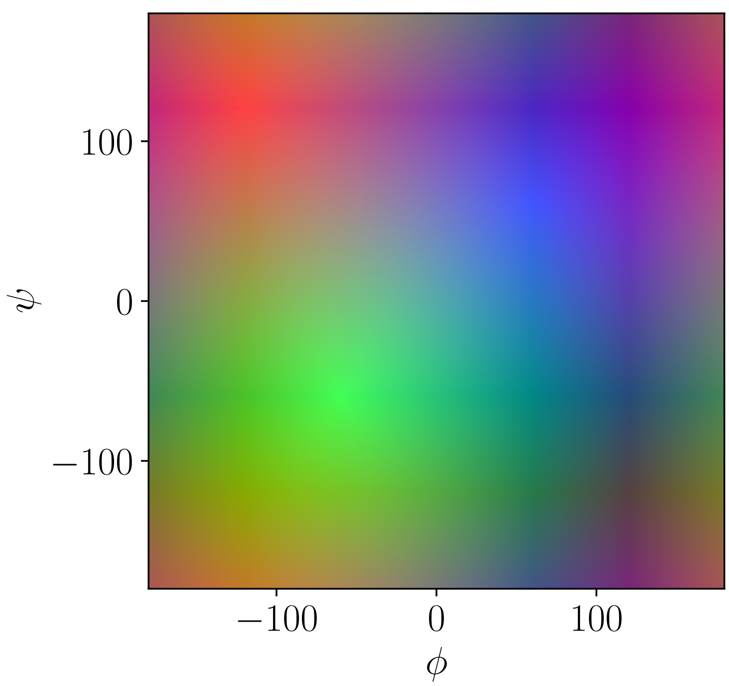

Figure S2: RGB values corresponding to phi-psi angle pairs used to assignment a unique color value to a dihedral conformation. The major dihedral states are indicated with more vibrant colors: red, green, and blue correspond respectively to β-strand, α_R-helical, and α_L-helical dihedral character. The code used to produce this mapping can be found at https://github.com/moldyn/ramacolor.[2,3]

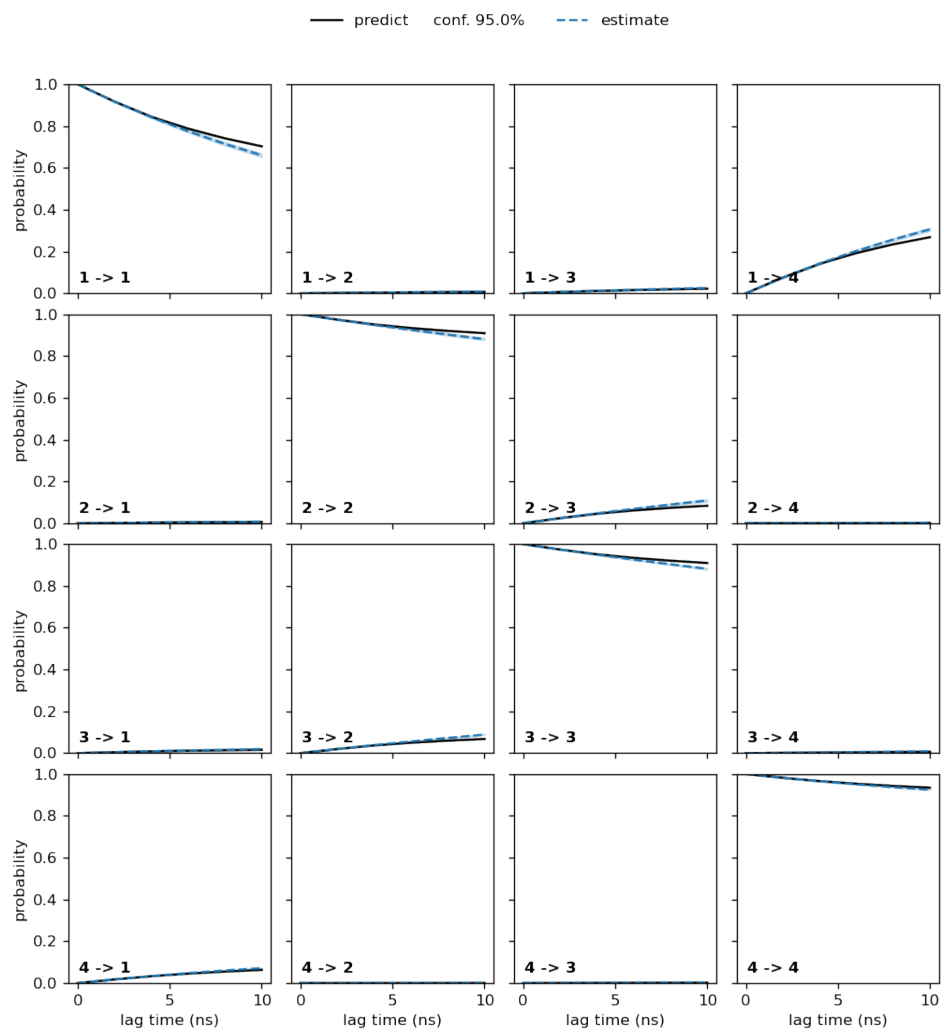

Figure S3: Chapman-Kolmogorov test for 4-state MSM resulting from weighted shape-GMM clustering on HP35. Computed using PyEMMA software.[1]

Table S1: Scores for a variety of clusterings of the HP35 trajectory. VAMP-2 scores were computed after construction of MSMs using a 10-fold cross validation on five trajectory segments using PyEMMA.[1] These were computed using either the first four ($k = 4$) or the first 10 ($k = 10$) eigenvalues of the transition matrix. Log likelihoods of a weighted shape-GMM (wSGMM) model were also computed on 10 randomly selected trajectory segments. A dynamic coring (here denoted dCore) procedure was applied to each clustering.[2] Error as estimated by 10-fold trajectory chunking are reported in parentheses.

Clustering Model	VAMP-2 Score (k=4)	VAMP-2 Score (k=10)	wSGMM Log Likelihood per Frame
wSGMM 4-state	3.3(2)		372.4(1)
wSGMM 4-state (dCore)	3.87(3)		372.0(1)
wSGMM 6-state	3.4(1)		379.5(1)
wSGMM 6-state (dCore)	3.89(1)		378.6(1)
uSGMM 6-state	2.9(1)		368.0(1)
uSGMM 6-state (dCore)	3.93(1)		367.66(9)
wSGMM 12-state	3.7(1)	7.2(6)	388.4(1)
wSGMM 12-state (dCore)[*]	3.91(1)	9.39(5)	383.16(9)
Stock 12-state (dCore)	3.91(1)	9.30(4)	364.2(3)
Stock 12-state (state specific dCore)[†]	3.93(1)	9.46(3)	363.9(2)

[*] Dynamic coring of wSGMM 12-state yielded a 10-state model.

References

(1) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. PyEMMA 2: A software package for estimation, validation, and analysis of Markov models. *J. Chem. Theor. Comput.* **2015**, *11*, 5525–5542.

(2) Nagel, D.; Weber, A.; Lickert, B.; Stock, G. Dynamical coring of Markov state models. *J. Chem. Phys.* **2019**, *150*.

(3) Sittel, F.; Stock, G. Robust Density-Based Clustering to Identify Metastable Conformational States of Proteins. *J. Chem. Theory Comput.* **2016**, *12*, 2426–2435.