Electronic Supplementary Information for Unsupervised learning of sequence-specific aggregation behavior for a model copolymer

Antonia Statt,[1] Devon C. Kleeblatt,[2] and Wesley F. Reinhart[2,3,a)]

[1)] *Materials Science and Engineering, Grainger College of Engineering, University of Illinois, Urbana-Champaign, IL 61801*

[2)] *Materials Science and Engineering, Pennsylvania State University, PA 16802*

[3)] *Institute for Computational and Data Sciences, Pennsylvania State University, PA 16802*

(Dated: 28 July 2021)

[a)]email:reinhart@psu.edu

I. LOCAL ENVIRONMENT UMAP

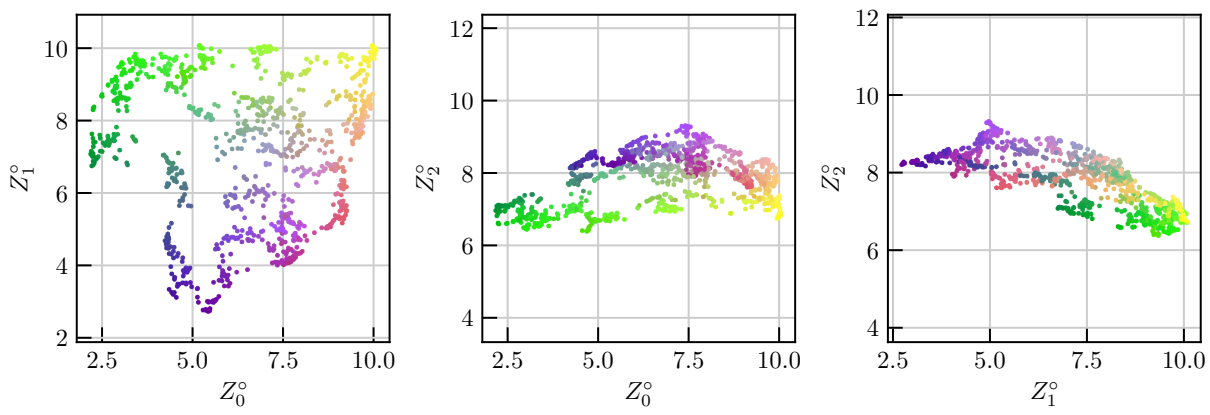

FIG. S1. Manifold obtained by UMAP for local (coarse-grained) environments, with HSV color scale for easy visualization in the simulation snapshots.

Fig. S1 shows the UMAP result for all local environments observed from the 1022 k-means sequences. There were 1000 environments from each sequence (2 per chain), taken from a single snapshot at the end of the equilibration by MD. The embedding was fixed for all other calculations of $\mathbf{H}^\circ$. A color scale was assigned based on a Hue-Saturation-Value mapping from the 3D coordinates in $\mathbf{Z}^\circ$. This color scale was applied to all particles shown in the main text, with detailed snapshots available in Fig. 2.

II. FINITE SIZE EFFECTS

In Fig. S2, we show morphologies obtained from 100 chains for all 1022 sequences determined by k-means. The 125 peripheral sequences are highlighted to show that these remain close to the periphery and are a reliable way to obtain the alpha-shape. This provides a 8× time savings compared to simulating all 1022 sequences. It is preferable to include only these peripheral sequences because they represent the morphological archetypes and most of the sequences in the center of the manifold are mixtures of those on the periphery.

In the main text, a new manifold was computed for the finite size study using only the peripheral sequences. This was important because UMAP does not provide robust placement of observations outside of the training domain. In essence, extrapolated points have no guarantee of topology preservation when projected into the learned low-dimensional

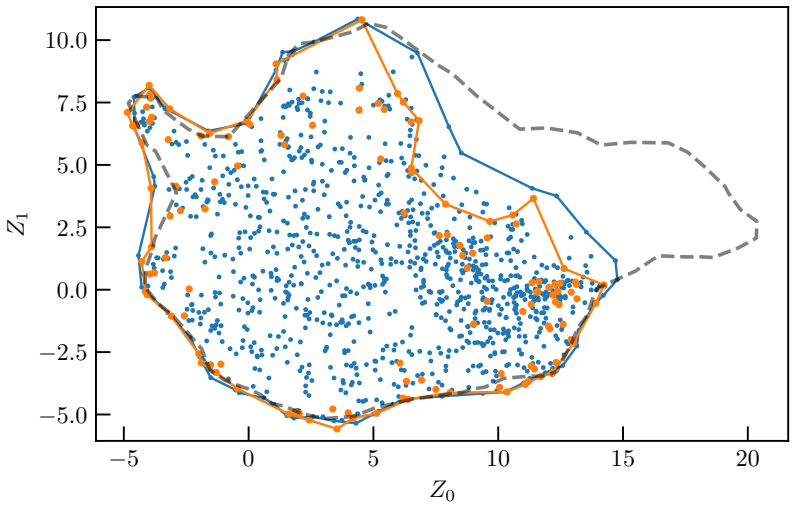

FIG. S2. Manifolds with $N = 100$ chains embedded into the k-means $N = 500$ manifold. Blue symbols are all 1022 $N = 100$ sequences, with blue line indicating the alpha-shape. Orange symbols are 125 peripheral $N = 100$ sequences, with orange line indicating the alpha-shape. Dashed black line corresponds to alpha-shape for all data at $N = 500$.

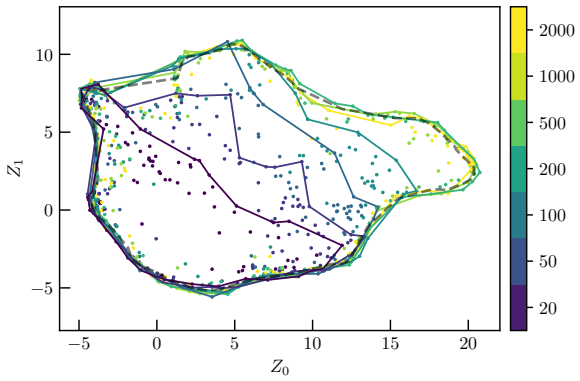

FIG. S3. Morphologies for $N = 20$, 50, 100, 200, 1000, and 2000 chains embedded in the k-means manifold from $N = 500$ chains. Symbols correspond to each observed periphery sequence. Lines correspond to alpha-shapes for each N (peripheral sequences only), with dashed black lines corresponding to alpha-shape for all data at $N = 500$.

space. Nevertheless, for ease of comparison we show the morphologies from the finite size study embedded into the $N = 500$ manifold here. As observed in Fig. 6, we see here that for low N, membranes, vesicles, and liquid droplets cannot form. For $N = 20$, wormlike

micelles also cannot form which is visible from the location of the purple points, although the alpha-shape does not capture the inward deflection at the left side. As N increases, the right edge of the alpha-shape moves up and to the right (towards vesicles and liquids), and eventually the full space is recovered. Note that unlike in Fig. 6, complete convergence appears to occur for $N = 500$, demonstrating the extrapolation problem described above.

III. DETAILS FROM THE EMBEDDING PROCEDURE

The main text gave a brief description of the acquisition of *features* and the procedure for *pooling* into *histograms* since this procedure follows closely the one introduced in our recent prior work[1]. Here we provide more details for representative examples to illustrate the procedure more explicitly for the macromolecular aggregates use case.

The histograms for each particle environment are shown in Fig. S4. Each set of three sub-panels corresponds to a single representative local environment. The raw features $\mathbf{F}$ are just an array and would be difficult to visualize, so this is after the first pooling step in Fig. 1. These histograms will be flattened into a single feature vector for embedding in the next step. The embedding occurs by comparing each of these feature vectors to each other and constructing a 3D vector in $\mathbf{Z}^\circ$ space, as shown in Fig. S1.

Once the feature vectors for each particle have been *pooled* and *embedded*, the collection of Z° vectors are further *pooled* into a single histogram representing the entire snapshot. This is indicated in Fig. 1 as $\mathbf{H}$. Examples from the structures in Fig. 2 of the main text are shown in Fig. S5 – note these are the actual histograms for the entire snapshot, whereas the features shown in Fig. S4 are representative of only a single particle in those snapshots. To complete the procedure, this $\mathbf{H}$ is finally *embedded* into a global structural space to yield a 2D vector which describes the global morphology.

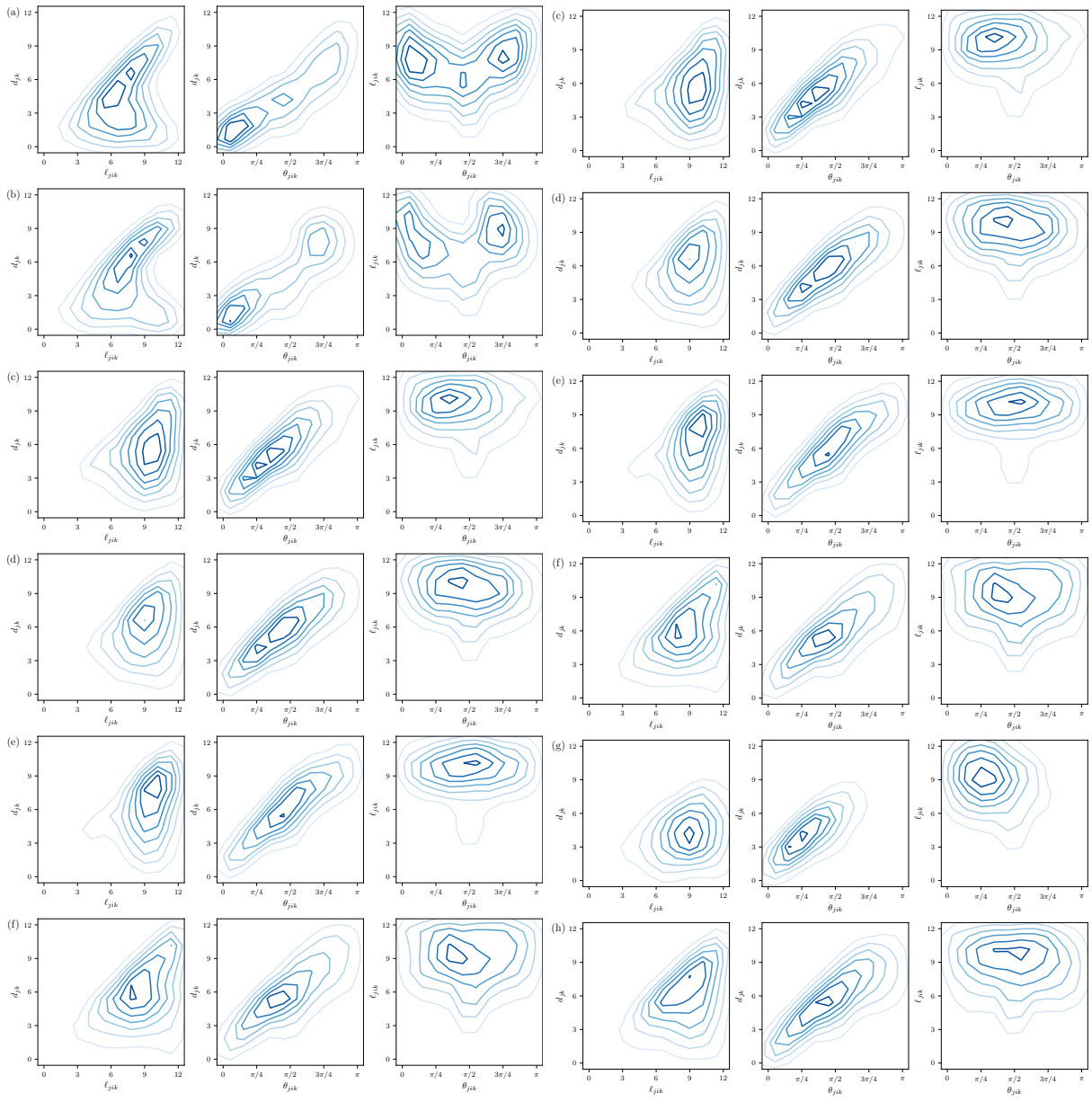

FIG. S4. Sample feature histograms (i.e., $\mathbf{H}^{\circ}$) corresponding to the structures in Fig. 2 in the main text. From each snapshot, a representative particle with the most common features (as determined by a kernel density estimation on the particle-level manifold) was selected. The three sub-panels for each structure represent different slices of the 3D histogram. Readers should refer to Ref.[1] for additional details and examples.

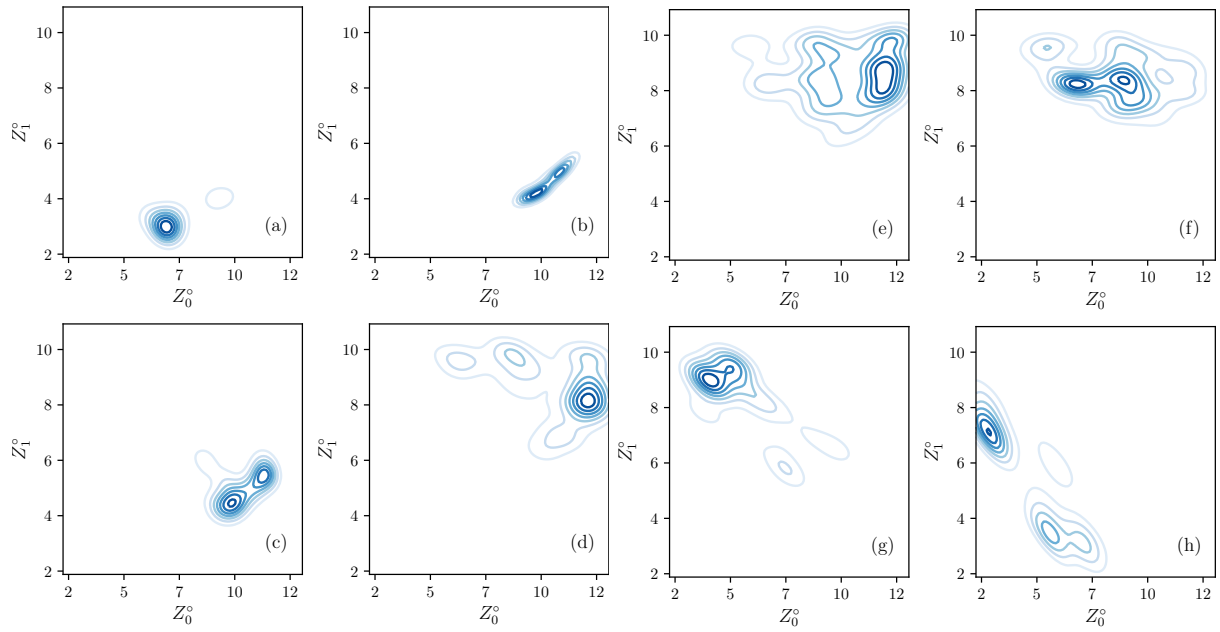

FIG. S5. Sample global features (i.e., **H**) corresponding to the structures in Fig. 2 in the main text.

IV. EXPERT-KNOWLEDGE-DERIVED ORDER PARAMETERS

We show many order parameters for global and single chain measures, motivated by Ref. 2 and a variety of sources from literature.[3–5]

- **genus**: genus of the surface (determined by Ovito, respects periodic boundary conditions, not normalized)
- **area**: area of the surface (determined by Ovito, respects periodic boundary conditions, not normalized)
- **volume**: volume of the surface (determined by Ovito, respects periodic boundary conditions, not normalized)
- **N clusters**: number of clusters in the system, cutoff $= 1.3\,\sigma$ for clustering
- **Rg**: average radius of gyration of each single chain
- **Ree**: average end-to-end distance of each single chain
- **A**: anisotropy value of each chain

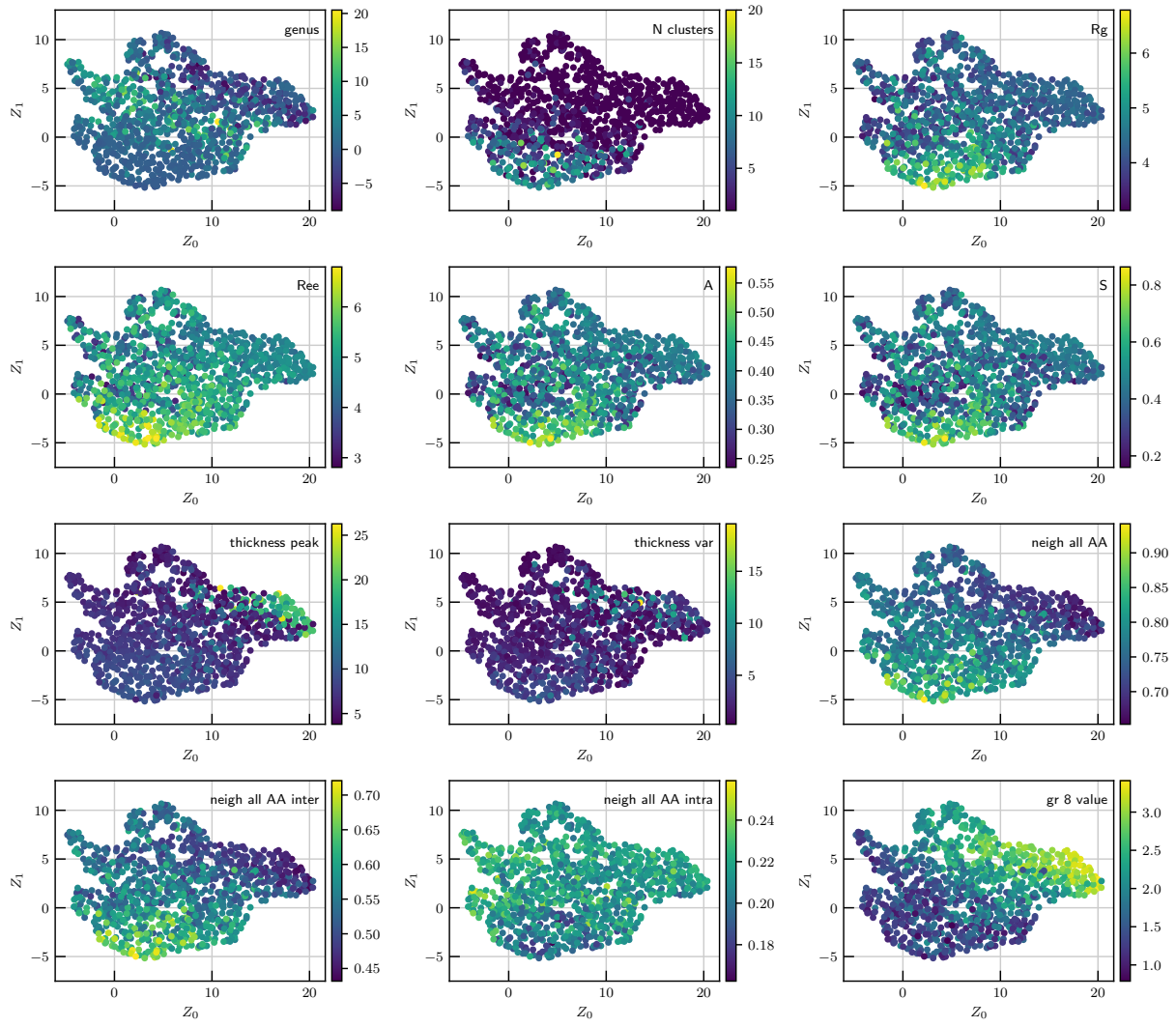

FIG. S6. Data for the k-means $N = 500$ manifold colored according to a variety of order parameters derived from conventional thermodynamic values.

- **S**: sphericity value of each chain
- **thickness peak**: peak location of the histogram of all thicknesses measured (obtained by fitting a Gaussian)
- **thickness var**: peak variance of the histogram of all thicknesses measured (obtained by fitting a Gaussian)
- **neigh all AA**: normalized number of A-neighbors in a $2\,\sigma$ radius around each A particle

- **neigh all AA inter**: normalized number of inter-chain A-neighbors in a $2\,\sigma$ radius around each A particle
- **neigh all AA intra**: normalized number of intra-chain A-neighbors in a $2\,\sigma$ radius around each A particle
- **gr 8 value**: value of the pair correlation function at $r = 8\,\sigma$

For **A** and **S**: A and S obey the inequalities $0 < A \leq 1$ and $-\frac{1}{4} < S \leq 2$, where

- $S < 0$: oblate
- $S > 0$: prolate
- $S = 2, A = 1$: rod shape
- $S = 0, A = 0$: perfect sphere
- $S = -\frac{1}{4}, A = 1$: pancake shape

For neighbor calculation:

- normalized by total number of neighbors
- directly bonded neighbors do not count (since they don't interact with LJ)
- should correlate well with LJ energy

For Ovito surface determination:

- surface, volume, genus, and area are determined for each cluster in the system, quantities are averaged over clusters
- for surface algorithm: smoothing level = 10, probe sphere radius is $2\,\sigma$ (using `ConstructSurfaceModifier.Method.AlphaShape`)
- for clustering: cutoff is $1.3\,\sigma$ (using `ClusterAnalysisModifier`)
- has unclear input parameters (probe sphere radius and smoothing level) cutoff for cluster is estimated from first non-bonded pair-correlation minimum

For thickness calculations:

- determine normal of all triangles from Ovito surface, determine distance to intersection with surface from normals
- fit Gaussian to most prominent peak in histogram of all intersection distances

REFERENCES

[1]W. F. Reinhart, "Unsupervised learning of atomic environments from simple features," Computational Materials Science **196**, 110511 (2021).

[2]A. Statt, H. Casademunt, C. P. Brangwynne, and A. Z. Panagiotopoulos, "Model for disordered proteins with strongly sequence-dependent liquid phase behavior," The Journal of Chemical Physics **152**, 075101 (2020), https://doi.org/10.1063/1.5141095.

[3]N. Rawat and P. Biswas, "Shape, flexibility and packing of proteins and nucleic acids in complexes," Physical Chemistry Chemical Physics **13**, 9632–9643 (2011).

[4]R. I. Dima and D. Thirumalai, "Asymmetry in the shapes of folded and denatured states of proteins," The Journal of Physical Chemistry B **108**, 6564–6570 (2004).

[5]J. Vymetal and J. Vondrasek, "Gyration-and inertia-tensor-based collective coordinates for metadynamics. application on the conformational behavior of polyalanine peptides and trp-cage folding," The Journal of Physical Chemistry A **115**, 11455–11465 (2011).