SUPPORTING INFORMATION

Nonlinear machine learning of patchy colloid self-assembly pathways and mechanisms

Andrew W. Long and Andrew L. Ferguson*

Department of Materials Science and Engineering, University of Illinois at Urbana-Champaign, Urbana IL 61801, USA

*Corresponding author. Tel: (217) 300-2354. Fax: (217) 333-2736. Email: alf@illinois.edu.

1. Modification of IsoRank graph matching algorithm

We define pairwise distances between clusters found in our trajectories by computing a measure of the similarity of the interaction graphs corresponding to each cluster using a modification of the IsoRank graph matching algorithm[1]. The original implementation of this algorithm computes similarity scores between the nodes of each graph as input to a greedy algorithm that successively pairs nodes between the two graphs possessing the maximum similarity scores. Alternatively, one could conceive of a more sophisticated approach that seeks to maximize the overall match score, as in the Hungarian algorithm[2]. These implementations work well when graph structures are sparse, but the highly connected nature of the cluster interaction graphs in this work cause these relatively simple assignment protocols to fail. Specifically, they frequently fail to find (near) optimal alignments between cluster pairs, and often have difficulty even mapping identical interactions graphs onto themselves. This motivated us to develop a modified IsoRank variant that preserves local connectivity in the greedy assignment process and results in improved alignment performance.

We start by finding the similarity scores for each node pairing between the two graphs as in the standard IsoRank algorithm and choose the pairing with the highest similarity score[1]. In the event that more than one node pair has the same score, one pairing is selected at random. The assignment process then proceeds by considering as candidates for the next pairing only those nodes in each graph that are connected to the most recently paired node in each graph that we refer to as the "central node". The pairing within this candidate set possessing the maximum similarity score is selected next. In the event of ties we select the node pair possessing the maximum degree of edges linking to previously assigned nodes, thereby favoring matches within the densest segments of each graph first. If ties still persist, one pairing is selected at random. Each node pairing made is added to a queue, and after the central node pair is exhausted of unassigned neighbors, the nodes in the pairing at the head of the queue are made the central nodes in each graph. This process is iteratively repeated until all node pairings are made.

Our modification drastically improves the performance of the matching for a set number of attempted greedy assignments as compared to a Hungarian algorithm approach, and we observe empirically superior and more robust alignment performance compared to the original IsoRank implementation. Nevertheless, we observe that our modified algorithm remains fundamentally greedy in nature, and as an NP-complete problem[1,3], guaranteed recovery of the optimal global alignment would require

exhaustive enumeration over all possible alignments. For clusters containing more than a handful of particles, this operation quickly becomes computationally intractable.

2. Determining Gaussian kernel bandwidth

Following Singer et al.[4], we determine an appropriate bandwidth of the Gaussian kernel used to soft-threshold the pairwise cluster distances matrix by constructing a log-log plot of the element sum of the soft-thresholded cluster pairwise distances matrix, $\mathbf{A}$, against the kernel bandwidth, ϵ. The plots for the tetrahedral patchy particles considering all 63,276 post-deislanding clusters identified in simulations at all eight temperatures are presented in Figure S1, and for the icosahedral patchy particles considering all 48,010 clusters identified in simulations at all nine temperatures in Figure S2. Each system exhibits a double-sigmoid character, indicative of diffusion processes at two different length scales. At kernel bandwidths below the intermediate plateau, the bandwidth is sufficiently large for each cluster to "see" neighbors within their own geometry class, which possess the same internal bonding structure, but different bond lengths. For ϵ values above the intermediate plateau, the bandwidth is sufficiently large for clusters to "see" clusters outside their own geometry class, and the data becomes globally connected. It is imperative that we select a bandwidth in the latter regime to assure that we model one global diffusion process, rather than an ensemble of disjoint processes, to synthesize a single unified diffusion map embedding of the data. Empirically, we find that selecting relatively large ϵ values within the upper plateau is necessary to attain global connectivity within our data, with insufficient kernel bandwidths producing highly disconnected diffusion map embeddings. For the case of tetrahedral patchy particles, we selected $\epsilon = \exp(10)$, and for the icosahedral particles $\epsilon = \exp(12.5)$.

3. Determination of embedding dimensionality

In Figures S3 and S4 we present the diffusion map eigenvalue spectrums for the tetrahedral and icosahedral patchy particle systems, respectively. We estimate the noise floor of the eigenvalue spectrum in each system using a permutation test to destroy the structure within the pairwise distances matrix by breaking correlations within the data[5]. Specifically, we randomly shuffled the elements of each row in the upper triangle of the pairwise distances matrix, $\mathbf{d}$, then reflected the upper triangle into

the lower triangle to form a shuffled, but symmetric, pairwise distances matrix, $\mathbf{d}'$. By randomizing the distances to the neighbors of each cluster, we have scrambled the structural relationships between clusters such that it is no longer necessarily true that if clusters i and j are structurally similar (i.e., $d_{ij} \approx 0$) that the structural distances of this pair of clusters to any other cluster k are also similar (i.e., $d_{ik} \approx d_{jk}$). (Mathematically, this shuffling procedure causes the structural distances to no longer behave as a metric space, since the triangle inequality can be violated (i.e., $d_{ik} \leq d_{ij} + d_{jk}$ may not be satisfied)). Effectively, we have scrambled the structural relationships between the points in the ensemble, and by applying diffusion maps to the scrambled pairwise distances matrix, we can estimate the largest eigenvalues one might expect to see for this system in the absence of structural correlations between the clusters in the ensemble. For each system, we computed the mean value of the largest non-trivial eigenvector of 10 independent shuffles of the pairwise distances matrix and adopted this value as an empirical approximation for the noise floor of the eigenvalue spectrum. The location of the floor is indicated by a horizontal broken line for the tetrahedral system at $(9.995 \pm 0.002) \times 10^{-4}$ in Figure S3, and for the icosahedral system at $(5.052 \pm 0.001) \times 10^{-4}$ in Figure S4.

To identify an appropriate embedding dimensionality (i.e., an appropriate value of k in Eq. 8), we utilize the L-method[6] to identify the first "knee" in the eigenvalue spectrum. For both the tetrahedral and icosahedral systems, the knee occurs at the third non-trivial eigenvalue, λ_4 implying an effective dimensionality of k=3. For the icosahedral system, λ_4 falls below the noise floor, leading us to define the spectral gap at the last non-trivial eigenvalue above the noise floor, λ_3, and motivating an embedding dimensionality of k=2. In neither case did we observe functional dependencies between the leading eigenvectors[7].

References

[1] Singh, R.; Xu, J.; Berger, B. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 12763–12768.

[2] Kuhn, H. W. *Naval Res. Logist. Quart.* **1955**, *2*, 83–97.

[3] Conte, D.; Foggia, P.; Sansone, C.; Vento, M. *Int. J. Patt. Recogn. Artif. Intell.* **2004**, *18*, 265–298.

[4] Coifman, R.; Shkolnisky, Y.; Sigworth, F.; Singer, A. *IEEE Trans. Image Process.* **2008**, *17*, 1891–1899.

[5] Ferguson, A. L.; Falkowska, E.; Walker, L. M.; Seaman, M. S.; Burton, D. R.; Chakraborty, A. K. *PLoS ONE* **2013**, *8*, e80562.

[6] Salvador, S.; Chan, P. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*, 2004; pp 576–584.

[7] Ferguson, A. L.; Panagiotopoulos, A. Z.; Debenedetti, P. G.; Kevrekidis, I. G. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 13597–13602.

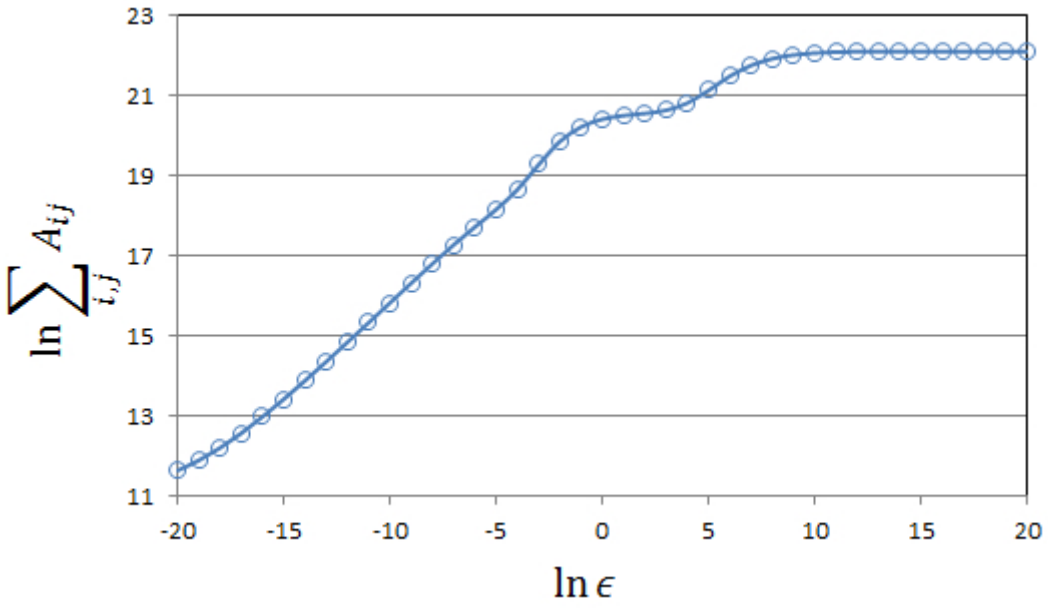

Figure S1: Log-log plot of the sum of the pairwise similarity matrix, $\sum_{i,j} A_{ij}$, as a function of soft-thresholding bandwidth used in Gaussian kernel for similarities computed between all realized clusters in tetrahedral simulations.

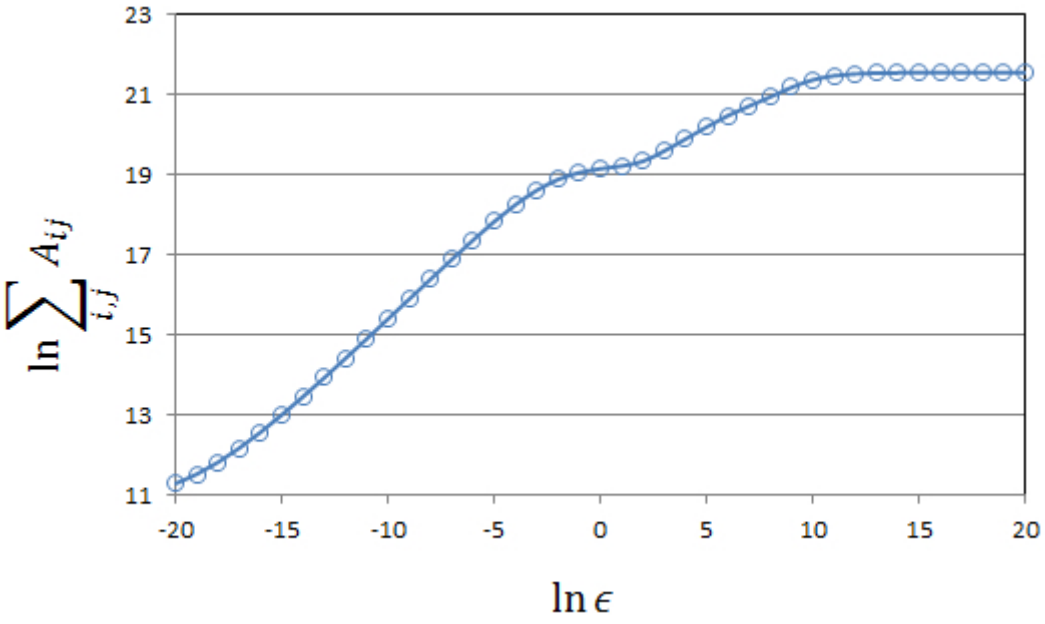

Figure S2: Log-log plot of the sum of the pairwise similarity matrix, $\sum_{i,j} A_{ij}$, as a function of soft-thresholding bandwidth used in Gaussian kernel for similarities computed between all realized clusters in icosahedral simulations.

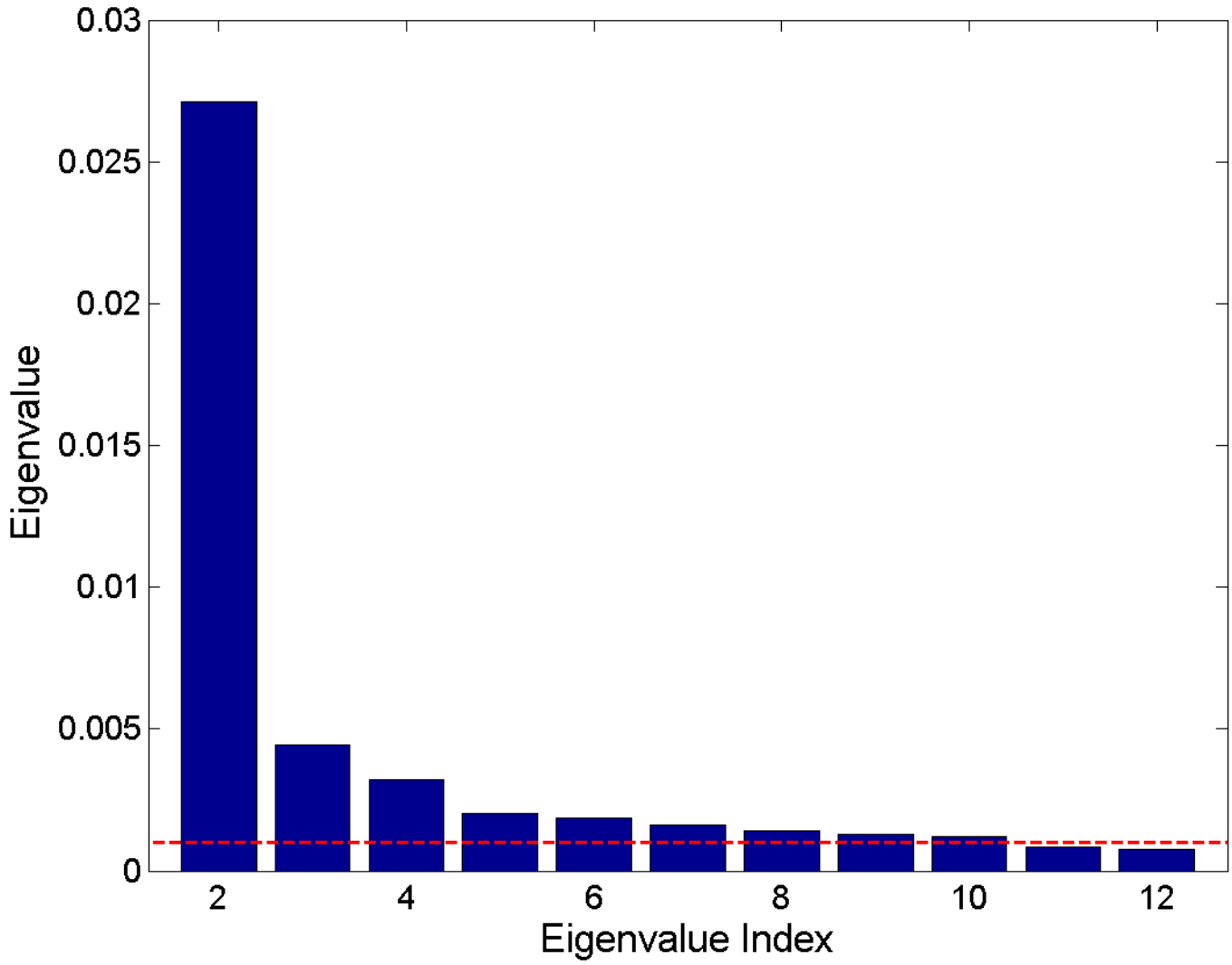

Figure S3: Diffusion map eigenvalue spectrum for the tetrahedral patchy particle system. The trivial unit eigenvalue, λ_1, has been omitted for clarity. The noise floor identified by a permutation test is indicated by a horizontal broken line. A spectral gap after the third non-trivial eigenvalue, λ_4, was identified using the L-method, motivating a k=3 dimensional diffusion map embedding of this system.

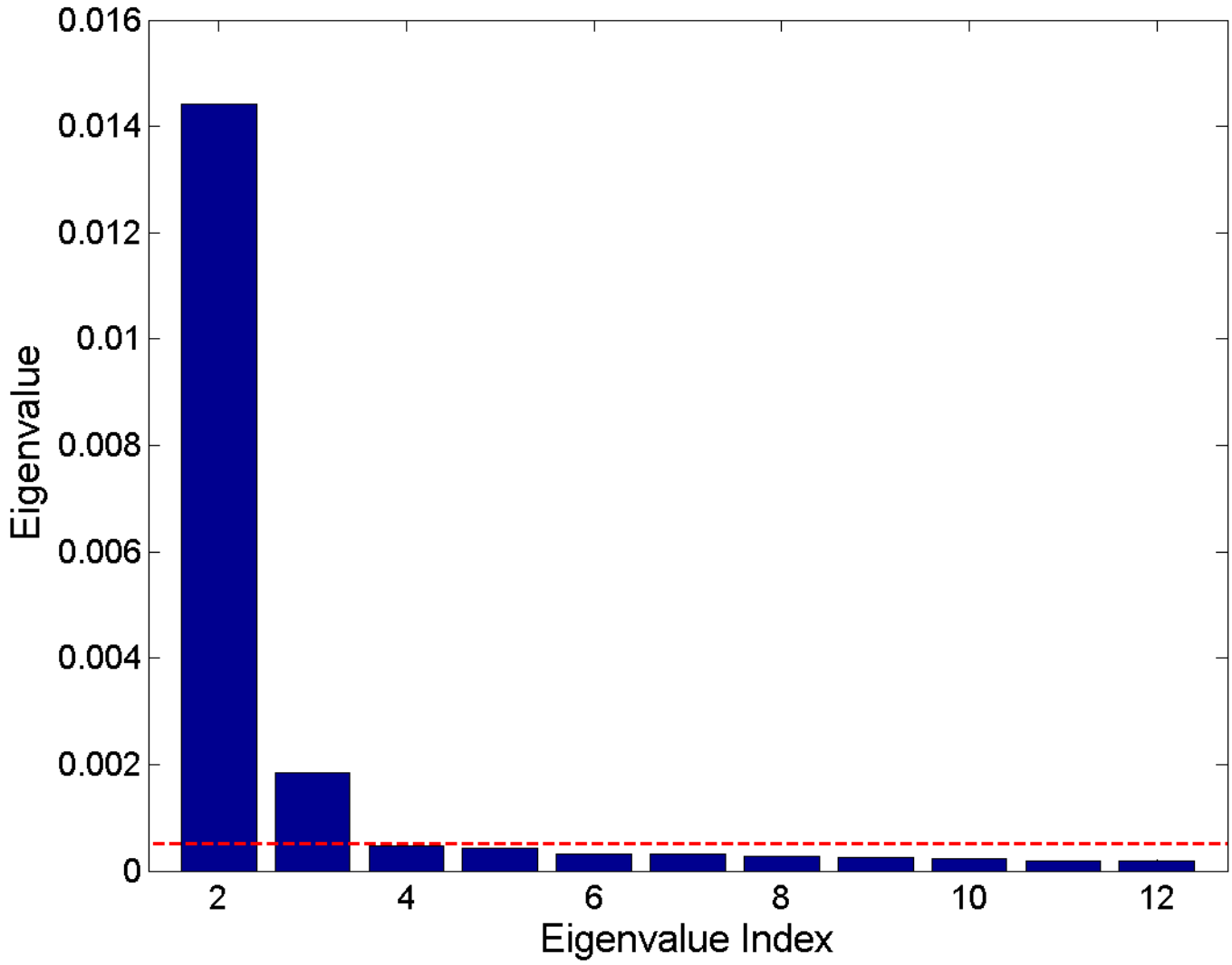

Figure S4: Diffusion map eigenvalue spectrum for the icosahedral patchy particle system. The trivial unit eigenvalue, λ_1, has been omitted for clarity. The noise floor identified by a permutation test is indicated by a horizontal broken line. A spectral gap after the third non-trivial eigenvalue, λ_4, was identified using the L-method. Since λ_4 resides below the noise floor, this led us to define the spectral gap at the last non-trivial eigenvalue above the noise floor, λ_3, motivating a k=2 dimensional diffusion map embedding of this system.