Supplementary Material: Comparing dimensionality reduction methods for local structural identification in colloidal systems

A. Ulugöl,[1] J.I. Bückmann,[1] R. Yang,[1] L.D. Hoitink,[1] A. van Blaaderen,[1] F. Smallenburg,[2] and L. Filion[1]
[1)] *Soft Condensed Matter and Biophysics Group, Debye Institute for Nanomaterials Science, Utrecht University, Princetonplein 1, Utrecht, 3584 CC, Netherlands*
[2)] *Université Paris-Saclay, CNRS, Laboratoire de Physique des Solides, 91405 Orsay, France*

(*Electronic mail: a.ulugol@uu.nl.)

(Dated: 18 December 2025)

S1. EFFECT OF ADDITIONAL HIDDEN LAYERS ON THE AUTOENCODER APPROACH

To assess whether the performance differences between autoencoders (AE) and UMAP reported in the main text arise from architectural under-capacity rather than intrinsic methodological differences, we systematically investigate how increasing AE complexity affects dimensionality-reduction quality.

Specifically, we study the performance of fully connected autoencoders as a function of network depth. An AE with n hidden layers consists of an encoder with n hidden layers followed by a decoder with n hidden layers. The number of hidden layers is varied from 1 to 16.

This analysis is performed on the supraparticle dataset, which represents the most challenging classification task considered in this work. For bulk crystalline structures, we find that even shallow autoencoders (two hidden layers) already achieve near-perfect separation, making them unsuitable for resolving performance differences at higher capacity.

The dataset is split into 60% training and 40% test data. Each particle is represented by a 13-dimensional descriptor vector, consisting of averaged bond-orientational order parameters and an additional scalar quantifying the displacement of the particle from the center of mass of its neighbors. Following Ref 1, all hidden layers are fixed to a width of 130 nodes, which is 10 times the number of descriptors. We use the hyperbolic tangent activation function throughout. Networks are trained using stochastic gradient descent with learning rate 5×10^{-4}, momentum 0.9, and weight decay 10^{-5}. Training is performed for 300 epochs.

We first examine the mean squared reconstruction error on the test set as a function of network depth. As shown in Fig. S1, the reconstruction loss decreases with increasing depth and plateaus beyond approximately 13 hidden layers. The shallow AE used in the main text (two hidden layers) achieves a test loss of approximately 0.20, while deeper architectures reach a minimum loss of approximately 0.15, corresponding to a reduction of about 25%.

This confirms that increasing model capacity improves reconstruction accuracy, but only up to a moderate depth, beyond which further gains are negligible.

To determine whether improved reconstruction accuracy translates into higher-quality low-dimensional representation, we evaluate the trustworthiness score[2], which quantifies how well local neighborhoods in the original high-dimensional space are preserved in the low-dimensional representation.

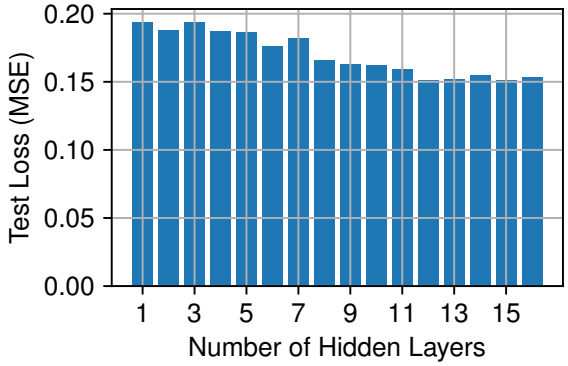

FIG. S1. Test-set reconstruction loss (mean squared error) of autoencoders as a function of the number of hidden layers in the encoder–decoder architecture.

The trustworthiness $T(k)$ is defined as

$$T(k) = 1 - \frac{2}{nk(2n-3k-1)} \sum_{i=1}^{n} \sum_{j \in \mathcal{N}_i^k} \max\left(0, m_{ij} - k\right), \quad \text{(S1)}$$

where $\mathcal{N}_i^k$ denotes the set of k nearest neighbors of point i in the low-dimensional representation, and m_{ij} is the rank of point j in the ordered list of neighbors of i in the original space. Trustworthiness values range from 0 (no neighborhood preservation) to 1 (perfect preservation).

We compute trustworthiness for $k \in \{10, 20, 40, 80, 160\}$, probing both local and increasingly global structure. The results are shown in Fig. S2.

Across all values of k, all AE architectures yield high trustworthiness scores of approximately 0.97. Increasing network depth leads to only marginal improvements: the maximum increase in trustworthiness across all k is approximately 0.005, occurring near 13 hidden layers. This gain is small compared to the overall magnitude of the metric.

For reference, we include trustworthiness scores obtained using UMAP with the same descriptor set and $n_{\mathrm{NN}} = 25$, as used throughout the main text. UMAP achieves the highest trustworthiness for $k \in \{10, 20, 40, 80\}$, outperforming all AE architectures in this regime. At k = 160, UMAP performs comparably to autoencoders with 6–7 hidden layers.

As expected, trustworthiness of UMAP decreases monotonically with increasing k reflecting the emphasis of UMAP on

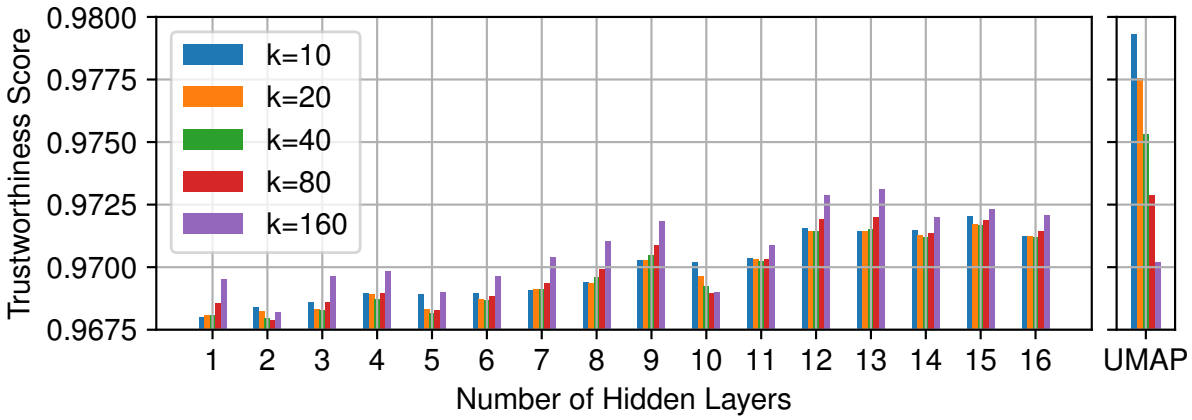

FIG. S2. Trustworthiness score of autoencoder dimensionality reduction as a function of the number of hidden layers in the encoder–decoder architecture, evaluated for different neighborhood sizes $k = 10, 20, 40, 80, 160$. Larger k probes increasingly global structure. The rightmost bars show the corresponding trustworthiness scores obtained using UMAP for comparison

preserving local structure at the scale defined by its neighborhood size.

These results demonstrate that substantially increasing AE capacity improves reconstruction loss but yields only minimal gains in neighborhood preservation. Within the scope of our work, deep autoencoders do not surpass UMAP in preserving local structure for the supraparticle dataset considered here. This supports the conclusions of the main text that UMAP offers a favorable balance between performance and complexity, while deeper autoencoders provide diminishing returns despite significantly increased architectural and training complexity.

S2. LOCAL STRUCTURE IDENTIFICATION USING A DEEPER AUTOENCODER

To further quantify the performance of deeper AEs, we select the AE with 13 hidden layers as a representative model. This architecture yields the highest trustworthiness score while also achieving (near-)minimal reconstruction loss.

We then apply Gaussian mixture model (GMM) clustering to the corresponding latent representations, considering models with up to 30 components. The resulting Akaike information criterion (AIC) and Bayesian information criterion (BIC) curves are shown in Fig. S3(a). Both criteria flatten beyond ~ 16 components, indicating diminishing returns when increasing the number of clusters. To retain some flexibility for the subsequent coarse-graining step, we select 18 components for the initial GMM fit.

Next, we perform entropy-based cluster merging. As shown in Fig. S3(b), the entropy exhibits two pronounced coarse-graining jumps at 4 and 7 clusters, analogous to the behavior observed for the AE baseline in the main text, and then transitions into a steadily increasing regime beyond 9 clusters. To resolve as many distinct structural classes as possible without over-fragmentation, we choose 9 clusters as the final cluster count. The resulting classification in the 3D latent space is

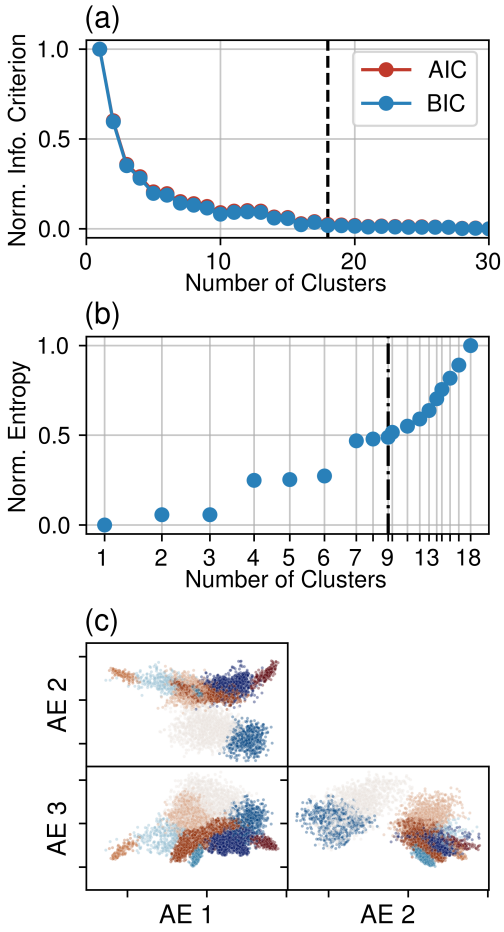

FIG. S3. Clustering of icosahedral supraparticles in the 13-layer-deep autoencoder projection. (a) AIC and BIC scores for GMM clustering. (b) Entropy as a function of the number of clusters after successive merging. (c) Projection and classification after entropy-based merging into 9 clusters.

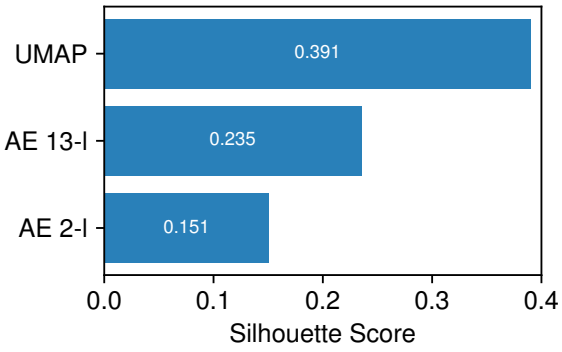

FIG. S4. Silhouette scores quantifying cluster separation for the final clustering from UMAP, 13-layer autoencoder (AE 13-l), and a 2-layer autoencoder (AE 2-l). Higher values indicate better separation between clusters in the low-dimensional projection. UMAP yields the highest silhouette score, followed by the deeper AE, while the shallow AE shows the weakest cluster separation.

visualized in Fig. S3(c).

To further quantify the separation of the resulting clusters, we compute the silhouette score, which measures the relative compactness of clusters compared to their mutual separation. The silhouette scores for UMAP, the 13-layer AE, and the 2-layer AE are shown in Fig. S4. UMAP yields the highest silhouette score, indicating the strongest separation between structural classes in the projected space. The 13-layer AE performs considerably better than the 2-layer AE, but remains below UMAP. This ranking is consistent with the trends observed in the trustworthiness analysis, indicating that increased AE depth leads to only modest gains in cluster separation relative to UMAP.

In Fig. S5 we visualize the clustering results for the same set of radial cuts of the same supraparticle shown in Fig. 18 of the main text. Each row corresponds to a different radial cut, while the columns show the results obtained using the 2-layer AE, the 13-layer AE, and UMAP, respectively.

Compared to the 2-layer AE, the 13-layer AE is able to resolve additional structural detail in the near-surface region. In particular, it distinguishes the first and second subsurface layers in a manner that is very similar to UMAP.

Differences between the methods appear in the fourth radial cut, where the 13-layer AE produces a more coarse-grained clustering than both the 2-layer AE and UMAP. Beyond the fourth radial cut, all three approaches yield nearly identical cluster assignments. The only notable exception is that the 13-layer AE does not resolve a cluster corresponding to the second layer of fivefold tubes, which is consistently identified by both the 2-layer AE and UMAP.

[1]E. Boattini, M. Dijkstra, and L. Filion, J. Chem. Phys. **151**, 154901 (2019).
[2]J. Venna and S. Kaski, in *International conference on artificial neural networks* (Springer, 2001) pp. 485–491.

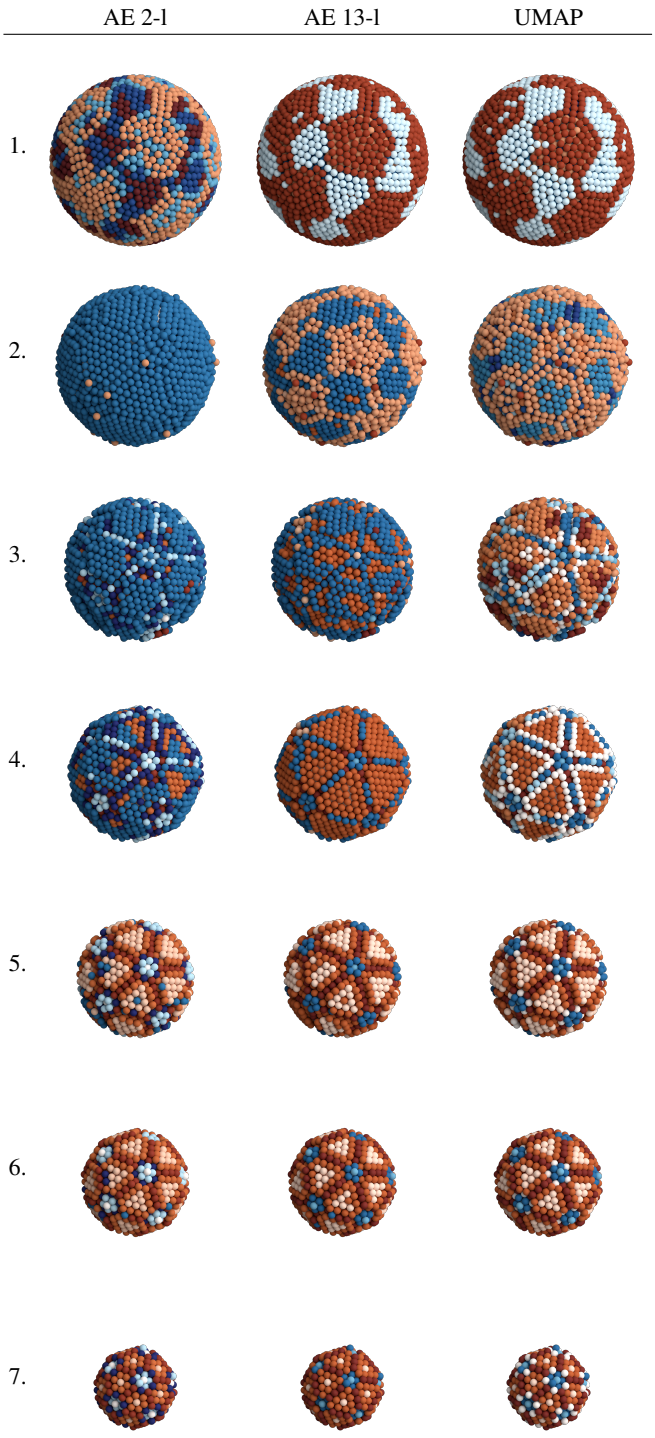

FIG. S5. Final cluster assignments for 13-layer-deep AE and UMAP across radial cuts of the same supraparticle configuration (rows).