Version of Record: https://www.sciencedirect.com/science/article/pii/S0927025624004294
Manuscript_7d98fbff5731816120a60f7710d41de9

COMODO: Configurable Morphology Distance Operator

Parth Desai[1] Namit Juneja[2] Varun Chandola[2] Jaroslaw Zola[2]
Olga Wodo[1]

[1]Materials Design and Innovation Department, University at Buffalo, Buffalo, NY, USA
{olgawodo}@buffalo.edu
[2]Computer Science and Engineering Department, University at Buffalo, Buffalo, NY, USA

Abstract

Data-driven approaches have been recognized as a new paradigm for establishing and exploring process-morphology-property relationships. However, typical exploration methods deliver high-dimensional morphologies that pose the challenge of extracting the key features and patterns that could guide the processing and materials design. The high dimensionality also hampers the organization of the data and the associated data analytics. As a solution, the currently available approaches either take a simplified view of the morphology, e.g., focusing on pixels in the morphology images, or apply transformations that average out structural descriptors of morphologies. To address these shortcomings, we propose a new computationally efficient and configurable distance operator that takes an intermediate approach. Our main idea is to represent the morphology as a graph where graph connectivity reflects the relative arrangement of components (e.g., grains, droplets) in the morphology, and the label of the graph vertices captures the domain-specific information of each characteristic domain. Next, given the graph abstraction, the distance between morphologies is computed using vectorized graph-based representation. Because both morphology graph structure and associated signature functions have clear interpretations, our distance measure can be easily tailored to specific applications. Our results demonstrate the superior performance of the proposed approach on data from simulation and synthetic data, including in real-world applications like morphologies clustering.

Keywords: morphology informatics, distance operator, graph, configurable signature functions, clustering.

1 Introduction

Methods of Machine Learning (ML) and Artificial Intelligence (AI) are becoming the fourth pillar of scientific discovery [14] complementing experimental, theoretical, and computational methods. The trend towards usage of ML and AI transforms the field of materials science and engineering with several prominent examples [13, 18]. Although machine learning tools are generic and can be used on most engineering problems, the efficacy and robustness of the associated models heavily rely on the expert knowledge captured in the model, data featurization, or associated semi-supervised or unsupervised transformations. The examples include reinforcing constraints in the convolutional neural networks-based models [20] to lower the demand on training data and featurization of the morphology to capture the aspects of the underlying physical processes [17]. Another example is the expert-defined descriptors that allow lowering the dimensionality of data for comparison of samples or materials [22]. However, distance or similarity measures are the most common customization. The distance measure is a central element of three classes of machine learning methods: clustering, neighborhood search, and indexing. Hence, by defining the distance measure reflecting the specifics of the task, many machine-learning methods can be leveraged at a minimal cost. This is also the scope of this work but in relation to the morphology datasets and related machine-learning tasks.

The materials morphologies play an important role in transport properties through the membranes [5], current generation in porous electrodes [12] or organic solar cells [22], mechanical properties of composite materials or fatigue of polycrystalline materials [3], to name a few. Because of the key role of morphology, establishing the morphology-property relationship of the materials is considered the holy grail of materials science. But other tasks like setting up and searching the databases, materials selection, and classification

are of key importance as well. Most of these tasks rely explicitly or implicitly on some form of distance or similarity measure and hence motivate this work. However, the distance measure is closely related to the morphology representation. The simplest form of morphology representation is a vector- (or matrix-) based representation of phase distribution that is typically readily available if micrographs are imaged by the microscope or are the results of numerical simulations. However, in such raw representation, the dimensionality is typically high (for two dimensions, at least 100^2 pixels, for three dimensions: 100^3 voxels), and some reduction is beneficial. Moreover, in such a raw format, the data may contain redundant or irrelevant information. This also demands some additional transformation to capture the critical information without redundancy. But most importantly for this work, the high dimensional data may cause the curse of dimensionality [2] hampering the meaningful distance calculation. Alternative representations are the statistical functions such as two-point correlation, which aims to capture the phase correlation within the morphology [6]. Such functions are typically translation-invariant and capture the average characteristics by operating in the frequency space. Compared to pixel-based representation, the dimensionality of such representation does not change, but with basic dimensionality reduction techniques, e.g., PCA [1], it can be successfully reduced. Nevertheless, the interpretability is reduced. Another method to represent the morphology is through a vector of expert-defined descriptors, such as volume fraction, average domain size, etc. Typically for a given morphology relatively small vector of physically meaningful descriptors is formed that captures averaged characteristics over the entire morphology. Although, in principle, a subset of these descriptors can be computed for each element of the morphology, such as grain or domain [4]; in practice, this may lead to an increase in the dimensionality of the mathematical representations, especially for very complex morphologies with a large number of grains or other elements. Nevertheless, with a vector of a small number of descriptors, the standard distance measured can be directly used. Finally, with recent advances in neural network models, latent space representation of autoencoders [19, 24] is being learned from the data in a supervised approach. However, these models require a large volume of data to find the latent representation of morphology that is, similar to statistical functions, not inherently interpretable. ‘

In the current work, we propose a morphology representation based on graph abstraction together with the configurable distance operator. Our representation balances the global and local morphological features. The connectivity of the entire morphology is captured by representing the connectivity of all components in the form of a graph, while local characteristics of individual components are captured through configurable signature functions. By defining one configurable signature function per basic component, we balance the dimensionality and interpretability. Using the representation, we define the associated distance operator that we call COMODO (COnfigurable MOrphology Distance Operator). To showcase the operator, we integrate it with the standard clustering algorithm from the machine learning domain. Through a series of clustering experiments, we demonstrate its configurability, resiliency to the size of the data, and ability to handle mixed types of morphology datasets.

2 Methodology

This section defines the proposed distance operator and the associated morphology representations. The section begins with the definition of three morphology representations: pixel-based, graph-based, and vectorized graph-based representation. The methods of conversion between the representations are also described. We close this section by formalizing the distance measure and how it is used to compute the distance matrix that is passed to the clustering algorithm.

In the pixel-based representation, the material morphology is binned into a uniform grid of pixels M that is enumerated by a 2-D vector. Each pixel captures information about the local state, e.g., phase. In this work, we consider a two-phase morphology, with the local state defined as:

$$M_{i,j} = \begin{cases} -1, & \text{if } (x_{i,j}, y_{i,j}) \in \text{ phase one} \\ 1, & \text{if } (x_{i,j}, y_{i,j}) \in \text{ phase two} \end{cases} \tag{1}$$

where $M_{i,j}$ corresponds to the local state at the location i, j in the input array M. The local state can take one of two values, $\{-1, 1\}$ to denote phase one or two, respectively. The size of this representation is $|M| = n_x \times n_y$, where n_x and n_y are the numbers of pixels in the x and y dimensions, respectively.

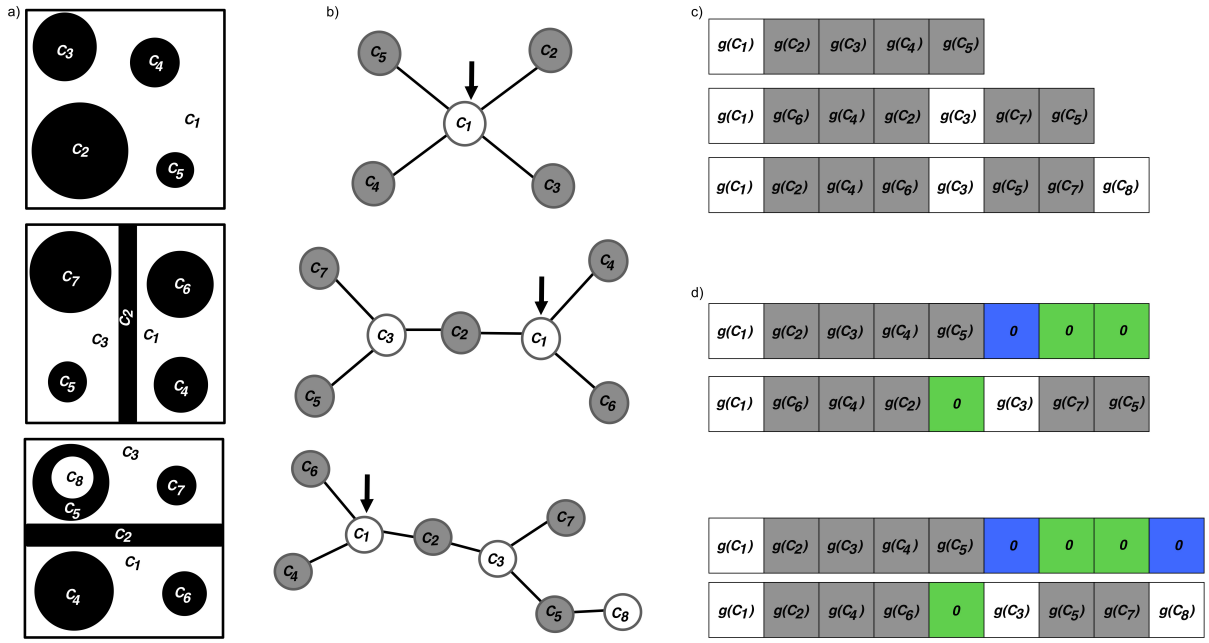


Figure 1: Three morphology representations used in this work: a) the pixel-based representation (black and white image), b) the graph-based representation, c) the vectorized graph-based representation, and d) the padded vectorized graph-based representation for two pairs of morphologies. In the first step, the pixel-based morphology is converted to a graph, where each vertex corresponds to the connected components in morphology - marked C_1 to $C_{N_{CC}}$ in panel a), where N_{CC} refers to the number of connected components. The graph is then vectorized, where each vector element stores the signature function value for the corresponding connected component. The order of elements in the vector reflects the structure of the graph (more details in the text). In panel d), two pairs of vectors are padded (the blue boxes indicate padding for extra white components while the green box indicates extra black components) to align the occurrence of the same phase elements in the vector and their length. Note that the vertices identified as the root vertices are indexed as C_1 and marked with the arrow in the panel b).

2.1 Graph-based morphology representation

In the graph-based representation, the morphology M is represented as a labeled undirected graph M^G [15]. A graph $M^G = (V, E)$ is defined as a set of vertices V and a set of edges E. Here, each vertex of the graph V corresponds to a connected component of a morphology. Formally, a connected component (C_i) is defined as a set of pixels that are of the same phase in the surrounding neighborhood. When all the pixels in the neighborhood are of the same phase, these pixels are considered members of the same connected component. The neighborhood of the pixel is defined as a set of pixels one pixel away.[1] The pixel neighborhood search algorithm [11, 23] is used to identify individual connected components (C_i) in the morphology M. Each vertex corresponds to the connected component in the morphology C_i, and it is labeled with the phase $p(C_i)$ and the value of the signature function $g(C_i)$. To capture both information, the aggregated label is used: $l(C_i) = p(C_i) \cdot g(C_i)$. To facilitate the padding of the vectors, the labels used are 1 and -1 for two considered phases.

The edge in the set E consists of a pair of vertices it connects. Formally, for a given morphology M and a set of vertices in V, two vertices $V' \in V$ and $V'' \in V$ are connected by an edge if and only if they are adjacent. Here, adjacent means that there exists at least one pair of pixels: (i', j') and (i'', j'') that are adjacent in the neighborhood of pixel-based representation. The pixel (i', j') originates from a set of pixels belonging to a connected component of V', and the pixel (i'', j'') originates from a set of pixels belonging to a connected component of V''.

The first two panels of Figure 1 illustrate three example morphologies in the pixel-based representation and the corresponding graph-based representation. Three morphologies in panel a) consist of a varying number of connected components: The first morphology consists of five connected components (C_1-C_5). The second morphology has seven connected components (C_1-C_7), and the third morphology has eight connected components (C_1-C_8). The difference in the number of connected components is paralleled by differences in their morphological features (e.g., size, shape) and arrangements (e.g., embedded or separating each other). The first morphology consists of one large white connected component that is adjacent to all four black connected components. As a consequence, in the graph-based representation - the first row of panel b) - the white vertex (C_1) is connected to all black vertices (C_2 to C_5). Compared to the first morphology, the second morphology contains one additional black-connected component, dividing the large white-connected component of the first morphology into two parts. As a consequence, for the second morphology, one additional black and one additional white connected component is identified. Also, the different graph structure is established. The first graph has a white vertex and four black vertices, representing white-connected components and four black-connected components in the pixel-based morphology. In the second graph, instead of one white connected component, the graph consists of two additional connected components, compared to the first graph the second graph consists of two white vertices (C_1 and C_3) that are connected to the black connected component (C_2). The addition of an extra connected component changes the structure of the graph, mimicking the change in the topology of the pixel-based morphology. For the third morphology, an additional white connected component (C_8) is embedded in a black connected component (C_5). Comparing it to the second morphology, one extra connected component is detected in the morphology, which also changes the structure of the graph. Here, the additional vertex in the graph is appended directly to vertex (C_5) because pixels belonging to this connected component are adjacent only to this connected component. Note, the orientation of the connected component does not affect the structure of the graph. For example, the black connected component (C_2) in the second morphology and the black connected component (C_2) in the third morphology divide the white component vertically and horizontally in respective morphologies, but the graph structure remains the same. Nevertheless, the anisotropy of the components can be captured by the signature function.

2.2 Vectorized graph morphology representation

Once the graphs are defined, the next step is to calculate the distance between them. The methods for exact graph comparison are computationally demanding as they require some additional function to establish an isomorphism between the vertices of the graphs under consideration. This problem, often referred to as graph alignment, is well studied, especially in the context of computational biology [9]. To overcome these challenges in the current work, we vectorize the graphs to compute the distance between vectors using standard distances like Euclidean or Cosine distance. Our vectorization aims to

[1] In two-dimensional problems, the neighborhood consists of first-order neighbors: north, south, east, and west and second-order neighbors: northeast, northwest, southeast, and southwest neighbors.

capture the hierarchy of the morphology structure, while the signature functions aim to capture the local characteristics of structural components.

The transformation from graph to vector begins with calculating the signature value of each vertex of the graph. We define three signature functions for reconfigurability:

- Surface to volume signature function:

$$g(C_i) = \frac{I(C_i)}{A(C_i)} \tag{2}$$

  where $I(C_i)$ is the perimeter of the component C_i and $A(C_i)$ is the area of the component C_i. This signature may be informative for materials properties dominated by diffusion and reaction. For example, when diffusion occurs in the bulk and reaction occurs at the surface, this signature function captures the propensity toward effective properties of each component in the morphology.

- Aspect ratio of each component C_i with the lengths of the component in the horizontal and vertical direction $L_x = L_x(C_i)$ and $L_y = L_y(C_i)$, are defined through two functions given below. This signature function may be informative for the mechanical properties of anisotropic materials.

  1) Normalized aspect ratio:

$$\begin{cases} g(C_i) = \frac{L_x}{L_y} \cdot (0.5 + (0.5 \cdot \tanh(\frac{L_x}{L_y} - 1))), & \text{if } (L_x/L_y < 1) \\ g(C) = 0.5 + (0.5 \cdot \tanh(\frac{L_x}{L_y} - 1)), & \text{otherwise} \end{cases} \tag{3}$$

  The signature function is defined such that for isotropic components, the function takes a value of 0.5, and for components elongated in horizontal and vertical directions, the function asymptotically reaches zero and one, respectively.

  2) Min-max aspect ratio:

$$g(C_i) = \frac{min(L_x, L_y)}{max(L_x, L_y)} \tag{4}$$

  which expresses the ratio of a longer length to a shorter length for a given component C_i.

Without loss of generality, other signature functions can be used, e.g., the fractal dimension, the Betti number, the curvature radii, etc.

Formally, the third representation M^V is an array of the aggregated labels for all the connected components C_i (vertices in the graph) in the morphology:

$$M^V = [g(C_1)p(C_1), g(C_2)p(C_2), ..., g(C_{N_{CC}})p(C_{N_{CC}})] \tag{5}$$

The order of the aggregated label reflects the graph structure and is established through a two-step process. In the first step, the vertex with the highest connectivity is identified. The connectivity of a vertex is determined by its degree (i.e., the number of nearest neighbors in the graph). The corresponding aggregated label is appended to the vector. In the second step, all its neighbors are ordered based on their values of the signature function, and the aggregated labels are appended to the vector M^V. In this work, we assume the descending order. The two-step process continues in an iterative fashion. Initially, the entire graph is searched for the vertex with the highest connectivity (and a priori selected phase).[2] Such vertex is selected as the root vertex, and its aggregated label initializes the vector.[3] At each subsequent iteration, only the nearest neighbors from the previous iteration are considered and appended to the vector. The described protocol follows the Breadth-First Search (BFS) algorithm [8] that is used to exhaust the search and establish the translation from the graph into the vector. The process concludes when all the vertices in the graph have been traversed.

To illustrate the vectorization process, panel c) of Figure 1 shows the vectors containing signature function values for three example morphologies. In this step, the graphs are transformed into their corresponding vector representations. For the first vector, the initial element of the vector corresponds to the highest connected vertex (root vertex) of the graph, which is C_1 (marked with the arrow in the middle panel). The successive elements are then added to the vector based on the value of the signature function of each vertex of the graph arranged in descending order. In the example shown here for the first morphology, the order of the vertices based on their signature function values is assumed to be

[2] In this work, we choose the white phase with the highest connectivity as a root vertex.

[3] The root vertex of a graph is the vertex all other vertices are derived from or connected to.

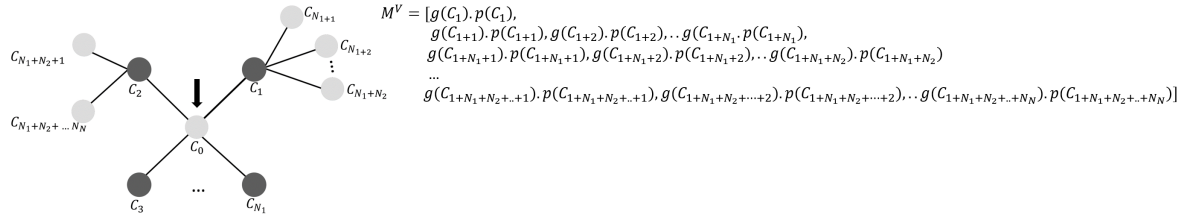


Figure 2: Example process of converting graph representation M^G to a vector representation M^V for complex morphologies. Note that indices N_1, N_2 up to N_N denotes the number of connected components considered at consecutive iterations of BFS algorithm. The total number of connected components: $N_{CC} = 1 + N_1 + N_2 + ... + N_N$.

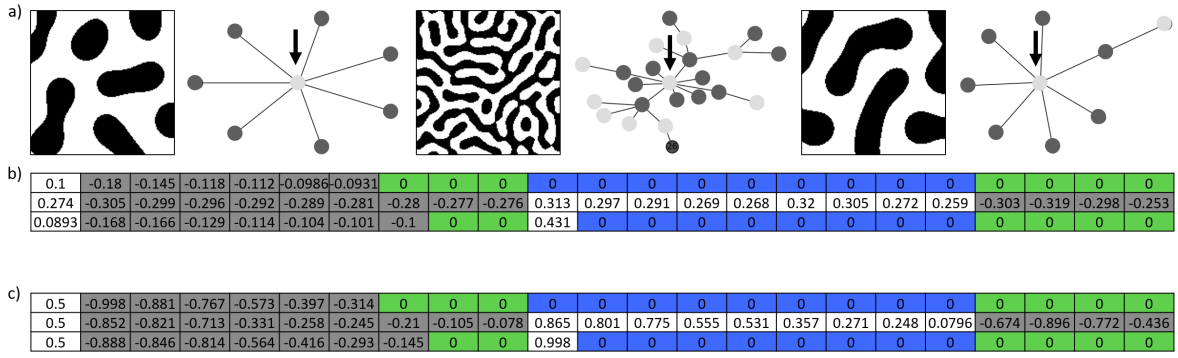


Figure 3: Three example morphologies and their corresponding graphs a) and corresponding vectors with aggregated labels for two signature functions: b) surface to volume) and c) aspect ratio. Each row in panels b) and c) corresponds to a vector M^V of an individual morphology M^V. Positive values correspond to white-connected components, and negative values correspond to black-connected components.

$g(C_2) > g(C_3) > g(C_4) > g(C_5)$, thus placing them in the order of their signature function following the first element. When multiple vertices have the same connectivity, like in the second morphology, we break the tie by choosing the node with a priori agreed color of component (here, the white node) or the node with the highest signature value.

2.3 COMODO distance measure

Once the graphs are vectorized, the standard distance measure between vectors can be used. Formally, consider a set $\mathcal{M} = \{\mathbf{M}_1, \ldots, \mathbf{M}_{N_m}\}$ of N_m morphologies, converted to graphs, vectorized and stored as a set $\mathcal{M}^V = \{\mathbf{M}_1^V, \ldots, \mathbf{M}_{N_m}^V\}$ using methodology introduced above. For all pairs of morphologies in the set, the standard distance between vectors is computed. In this work, we include the results for the Euclidean distance between two vectorized and padded graph-based representations.[4]

Depending on the task at hand, all-to-all distances may be calculated, or morphologies may be compared in the group. For the former, the distances are computed for all pairs of morphologies in the input set and form the distance matrix D. A distance matrix D is a square matrix of size $N_m \times N_m$ representing the pairwise distances between morphologies in a given set.

When the analysis is performed on the dataset, the vectors of individual morphologies may be of different lengths as corresponding morphologies may consist of a varying number of connected components. To calculate the distance between any vectors using standard methods, the vectors must be padded to the same length. The pairs of vectors are padded to align the first occurrence of the element corresponding to a given phase (here, black or white connected component/vertex) and then add the neutral elements (here, zero) in the empty elements.

[4]Cosine distance are also tested, and results are consistent with those for Euclidean distance.

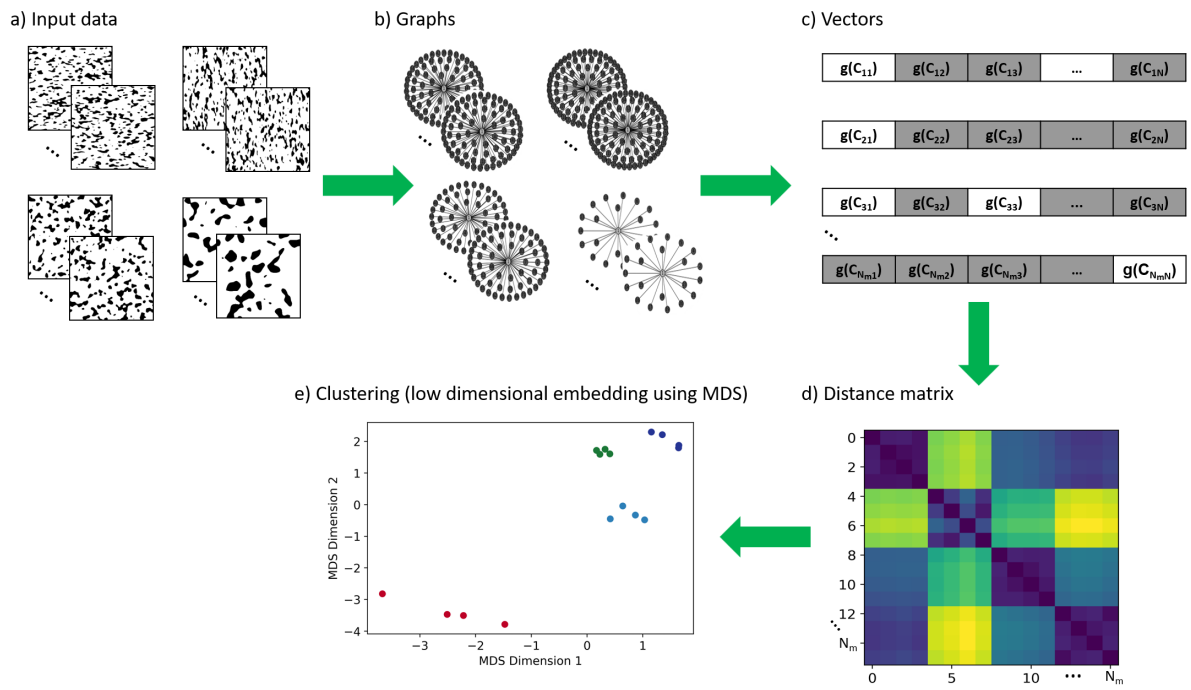


Figure 4: The clustering workflow using COMODO: The raw data is transformed into a graph, which is later converted into a vector. The distance matrix is then computed by calculating the Euclidean distance between vectors. For visualization purpose, MDS is used to project the morphological data into low-dimension space, and DBSCAN is used for clustering of the data

Panel d) in Figure 1 shows two examples of vector padding. In the first example shown here, the first vector has fewer elements that include four black elements and one white connected component. The second vector has one less element corresponding to the black-connected component. Hence, one additional neutral element is added to the second vector just after the fourth element, and it is marked green. Also, to pad for the additional elements in the second vector, three additional neutral elements are added to the first vector at the end. All these elements are identical and are only marked with colors to visually highlight the target padded element. The neutral elements padding for the black phase are color-coded green, while the padding of the neutral element for the white phase is color-coded blue. Similarly, for the second pair of vectors, the first vector is padded with four neutral elements at the end, while the second vector is padded with one neutral vector at the fifth element. The same coloring is kept for the visualization.

To close this subsection, Figure 3 provides examples of three morphologies used in this paper with the values of graphs annotated with the aggregated labels that are vectorized and padded. Panel a) in Figure 3 shows the three morphologies and their graphs; the other two panels show the corresponding vectors for two signature functions: the surface-to-volume (panel b) and the normalized aspect ratio (panel c). Each row represents the vector of the individual morphology. Note that vectors are padded with respect to the longest vector. The three morphologies are diverse, which is mirrored in the different graph structures, both in terms of the number of vertices and the connectivity of vertices in the graphs. The second morphology has the highest number of connected components and the longest corresponding vector. The vectors of the first and third morphology are padded to match the length and the connectivity pattern of the second morphology. Similarly to previous examples, for visualization, the white neutral elements are marked in blue boxes, and the black neutral elements are marked in green boxes. The elements with positive values correspond to the white components, while elements with negative values correspond to the black components. Note that when the signature function changes, the values change. The range of the aspect ratio function is larger, with higher absolute values capturing elongated domains.

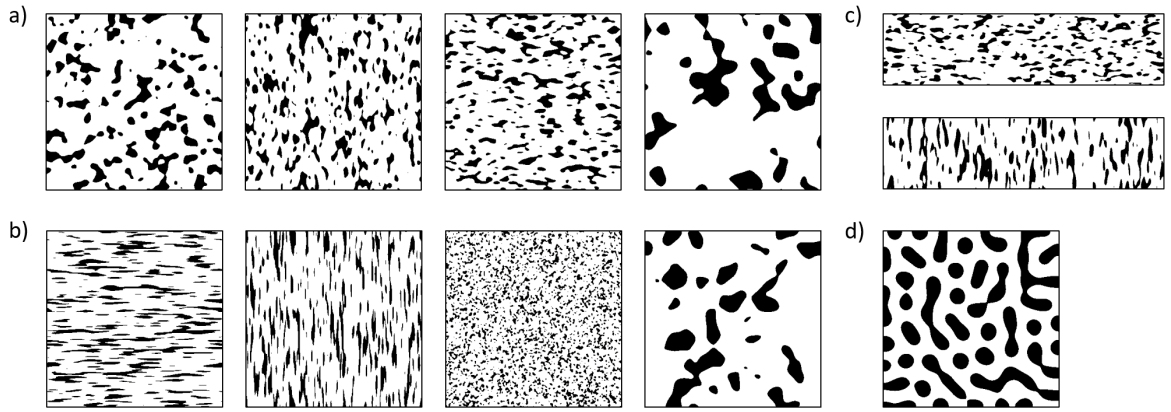


Figure 5: Representative morphologies used to evaluate COMODO performance: a) square domain composite data with grain size (20,40), (40,20), (40,40), (80,80); b) square domain composite data with grain size (10,10), (10,80), (80,10), (80,80); (c) thin film composite data with grain size (20,40), (80,20); (d) square domain spinodal decomposition data with $\phi = 0.54$ and $\chi = 2.2$.

2.4 Workflow

Defined above distance operator can be integrated with any machine learning pipeline that explicitly relies on the distance matrix, like clustering and classification, among many other techniques. In the current work, we integrate it with the clustering workflow and showcase three morphology datasets to demonstrate its configurability and resiliency to the size of the dataset and to the diversity of the datasets. The workflow steps are schematically shown in Figure 4. The input to the workflow is the set of morphologies in the pixel-based representation. The input data is transformed into graphs as shown in panel b) of Figure 4. After creating the graphs and selecting the signature function, the graphs are transformed into corresponding vectors as shown in panel c) of Figure 4. For a set of vectorized morphologies (as detailed in the above subsections), an all-to-all distance matrix is computed. Given the distance matrix (panel d), clustering is performed. To find how the data is separated into different clusters (panel e), DBSCAN is used [10].

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that uses a distance matrix to group data points together based on point density. DBSCAN defines clusters as areas of high density separated by areas of low density. It identifies outliers as noise points that do not belong to any cluster. In general, the method uses two hyper-parameters: minimum number of points and cutoff distance. The points are separated as core points (points in the cluster that have the minimum number of points in the specified cutoff distance) and border points (points that have at least one core point at the cutoff distance). The noise points are identified as points that are neither a core point nor a border point. The choice of the hyperparameters affects the accuracy of the results, and parameter tuning is required. See Figure S2 in Supplementary Information for the effect of the cutoff distance (hyper-parameter) selection on the accuracy of the clustering.

Finally, for visualization, multi-dimensional scaling (MDS) is used to project the data from high dimensional space into a low dimension subspace. MDS is a dimensionality reduction technique [16]. MDS translates the distance matrix between data points into the coordinates in the low dimensional subspace such that the proximity of individual points in the original space (provided in the distance matrix) and the corresponding distances are preserved in the low dimensional subspace. MDS is one of many visualization techniques, and in this work, we leverage it to visually display the clustering results.

3 Results and discussion

3.1 Data generation

To evaluate the distance measure proposed in this work, two types of data are used: the morphology typical for the spinodal decomposition and the composite type of morphology. Example morphologies are depicted in Figure 5. Panels a) to c) illustrate representative morphologies of composite data with a different aspect ratio of grains. While panel d) displays the representative morphology of spinodal data.

The composite type of morphologies is generated using an open-source generator [7]. The generator uses Gaussian random fields to create morphology and then the filter-based approach to create statistical replicas. The capability to generate statistical replicas sharing some morphological features is an important feature of the generator. In this work, we use this generator to create replicas of morphologies by specifying the average grain size in two major directions, L_x and L_y, while fixing the volume fraction.

The spinodal decomposition data is generated using the Cahn-Hilliard equation solver [21] to model mixing in the binary system of immiscible components. Similar to the composite data generator, the solver can be setup to generate statistical replicas (by varying initial noise field). This solver is used to generate mixed type of morphology with elongated and isolated domains by setting $\phi = 0.54$ and keeping the interaction parameter χ fixed. More details on the model can be found in our prior work [21].

Using two generators, this work uses different datasets organized into the following cases:

- Composite moderate aspect ratio dataset with varying grain size aspect ratio: This dataset consists of morphologies with the following grain size settings (L_x, L_y): (20,40), (40,20), (40,40), (80,80). Example morphologies are depicted in panel a) of Figure 5. Note two pairs of settings with the same aspect ratios are chosen. The first two settings have the same aspect ratio with switched grain lengths for the major directions, while the last two settings have re-scaled grain size lengths while keeping the same aspect ratio. This is the ideal setting to test the signature functions of COMODO. Small and large datasets are generated (i) small dataset of 400 morphologies with 100 replicas per grain size setting and (ii) large dataset of 4,000 morphologies with 1,000 replicas per grain size setting.

- Composite extreme aspect ratio dataset also with varying aspect ratio but two extreme configurations: The morphologies generated consisted of grain sizes of (L_x, L_y): (10,10), (10,80), (80,10), (80,80). Example morphologies are depicted in panel b) of Figure 5.

- Mixed morphology dataset with diverse morphology types: This dataset consists of morphologies with various sizes of the domain in composite morphologies, the type of morphology (composite and spinodal decomposition), and the size of morphology (square and thin films). The following configurations are used in the dataset: (i) Composite square morphologies with grain sizes (10,80) and (80,80). (ii) Thin film morphologies with grain sizes (20,40), (80,20). (iii) Spinodal decomposition square morphologies ($\phi = 0.54$, $\chi = 2.2$). This dataset consists of 500 morphologies with 100 replicas per configuration.

Two morphology sizes are generated here: square morphology and thin films. The size for the square morphologies is 400×400 pixels (n_x,n_y), while for the thin film: 200×800 pixels. Note that the size is selected to keep the consistent size of 160 thousand pixels per morphology. The first two cases use square morphologies, while the third case uses a mix of square and thin film morphologies. For consistency, the grain sizes are provided in the units of pixels.

In the following section, the distance operator performance is evaluated for different datasets generated. In the first set of experiments, the configurability of COMODO is showcased using the two signature functions. The next study demonstrates COMODO capability to handle large sizes of data. Finally, the third experiment demonstrates how COMODO can distinguish between different types of morphologies in a dataset.

In all experiments, the DBSCAN algorithm is used to cluster the data. The cutoff distance for DBSCAN is chosen based on the Rand Index. The Rand Index is a measure that quantifies the similarity between two data labels, here between the true labels and labels from the clustering. Rand Index provides a single numerical value as an evaluation metric (more details are provided in the Supplementary Information). In this work, the statistical replicas sharing the same morphological features and generation settings are considered to share the same true label.

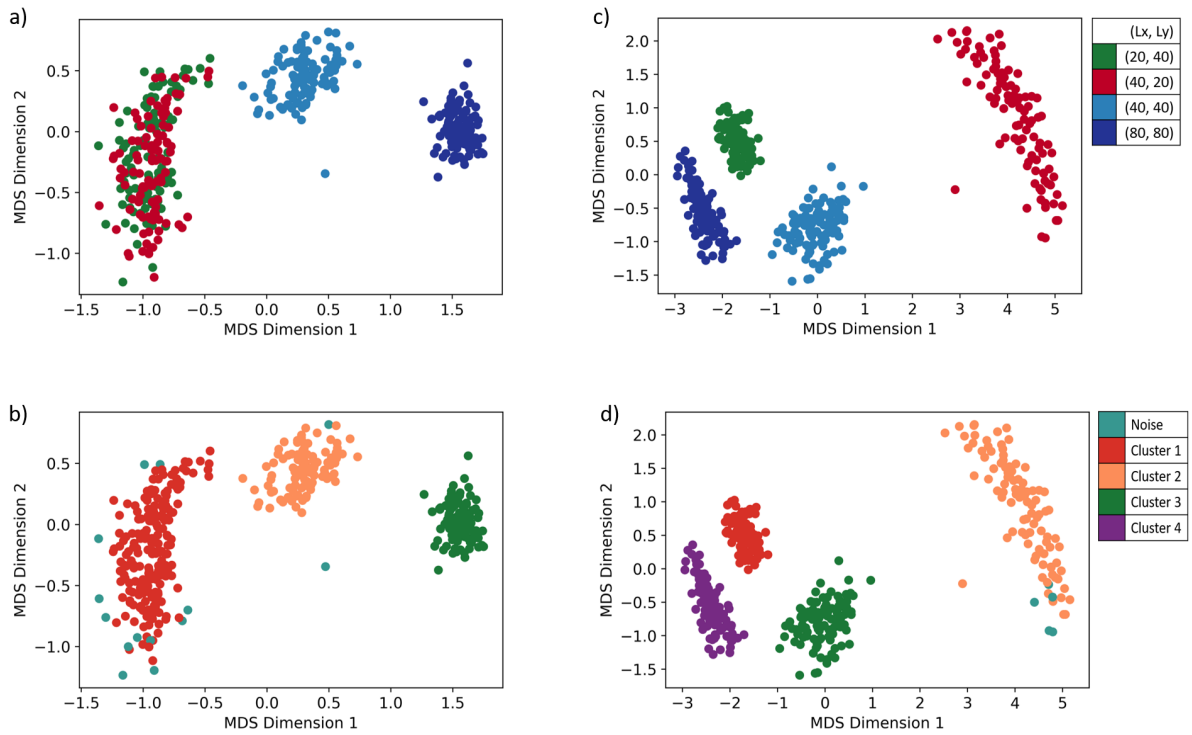


Figure 6: Clustering results for two signature functions on **moderate aspect ratio dataset**: a) low dimensional projection of data using **surface to volume** signature function where points are color-coded using the generation settings, b) the same projection of data as in above panel but color-coded with the indices of clusters (**surface to volume** signature function used to cluster data), c) low dimensional projection using for the signature function of **aspect ratio** where points are color coded with the generation setting, d) the same projection as in panel above but points are color coded with indices of clusters (**aspect ratio** signature function used to cluster data).

3.2 Distance measure is configurable

For the first experiment, we use two datasets: the composite moderate aspect ratio and the composite extreme aspect dataset, as defined in the data generation subsection. As a reminder, the grain sizes have an aspect ratio of (L_x, L_y) [(20,40), (40,20), (40,40), (80,80)], with 100 replicas for each grain size. Similarly, the extreme aspect ratio dataset also consists of 400 morphologies with grain sizes of (L_x, L_y) [(10,10), (10,80), (80,10), (80,80)] with also 100 replicas for each grain size setting. For both datasets, the morphologies are of the same size 400×400 (square morphologies) in pixel-based representation.

The top row on Figure 6 shows the true labels of the composite moderate aspect ratio dataset, and the bottom row depicts the clustering results for two signature functions. All results are depicted in the low dimensional subspace established using MDS, with each point representing one morphology. The left column depicts the results for the signature function of surface to volume, and the right column includes the results from the normalized aspect ratio signature function.

We begin the analysis with the results for the surface-to-volume signature function. The capabilities of COMODO are assessed in two ways: qualitatively by inspecting the distribution of points in the low dimensional subspace obtained through MDS and quantitatively by analyzing clustering results from DBSCAN, including Rand Index values. Starting with a qualitative assessment and panel a) of Figure 6. The set of 200 morphologies with the average grain size (20,40) and (40,20) are marked with the colors green and red to denote the different settings of morphology generation. They are grouped together when projected to the low-dimensional subspace, because these morphologies share a similar surface-to-volume ratio – the characteristics captured through the signature function. As intended, with this signature function COMODO operator ignores the anisotropy of the grains while focusing on the surface-to-volume ratio of constituting grains. Consequently, the distances between morphologies with these two settings are short. MDS preserves these distances and projects the points to similar regions in the low dimensional subspace - as depicted in panels a) and b). Two remaining groups of morphologies, marked light and dark blue in panel a), are located in different subspace regions. These morphologies consist of isotropic grains of sizes (40,40) and (80,80), respectively. Although this pair shares the aspect ratio of the grain sizes, the surface-to-volume differs between them. Additionally, morphologies differ in terms of the number of connected components per morphology and, hence, their graph structure differs. When grains are larger, fewer vertices in the graphs are identified. Consequently, the vectors are shorter, which leads to longer distances between morphologies (40,40) and (80,80). COMODO captures these unique characteristics of morphologies and when combined with MDS, morphologies with grain size $(40, 40)$ and $(80, 80)$ are projected to different locations in the low dimensional subspace.

However, MDS projections provide only qualitative insight into the data. Hence, we also include the clustering results for quantitative insight - see panel b) of Figure 6. The morphologies are color-coded based on the cluster index from the clustering - DBSCAN method. DBSCAN identifies the morphologies with grain sizes $(40, 20)$ and $(20, 40)$ as highly similar, as expected, and the other two types as belonging to separate clusters. These results mimic the true labels with the maximum value of the Rand Index of 0.92 (see panel a) in Figure S2 of Supplementary Information). When calculating the Rand Index, for this signature function, morphologies with grain size (20,40) and (40,20) share the same true label. Only a few morphologies are determined as noise points - points marked with a light teal color in the figure.

We now proceed to discuss the results for the second signature function: normalized aspect ratio. Panels c) and d) of Figure 6 include the relevant plots for the same dataset: moderate aspect ratio. Panel c) depicts the distribution of the morphologies in low dimensional subspace using the same legend as panel a). With this signature function, MDS projects the data into four groups of points, mimicking the data generation scheme. Similarly, DBSCAN also clusters the data into four clusters as shown in panel d) of Figure 6. As expected, morphologies are organized as clusters 1 and 2 (orange and red in panel d)) as they have similar graph structures but different values of the signature function; hence, they are recognized as distinct. This behavior is different from the results of surface-to-volume signature function (panels a) and b) of Figure 6), as normalized aspect ratio signature function recognizes morphologies with different aspect ratio: $(20, 40)$ and $(40, 20)$ as different. At the same time, morphologies with similar grain aspect ratios, $(40, 40)$ and $(80, 80)$, are also recognized as distinct, as although they share the same aspect ratio, the graph structure differs in terms of the number of components. The above results demonstrate the importance of signature function selection but also the robustness of the distance measure to capture the key characteristics of datasets.

The third dataset studies the morphologies with the extreme aspect ratio. Figure 7 summarizes the results for the composite dataset with the extreme ratio of grains. Similar to previous plots, the results in the low dimensional projection in panel a) show similar trends to the clustering results of the moderate aspect ratio dataset. For the surface-to-volume signature function, morphologies with grain sizes (10,80)

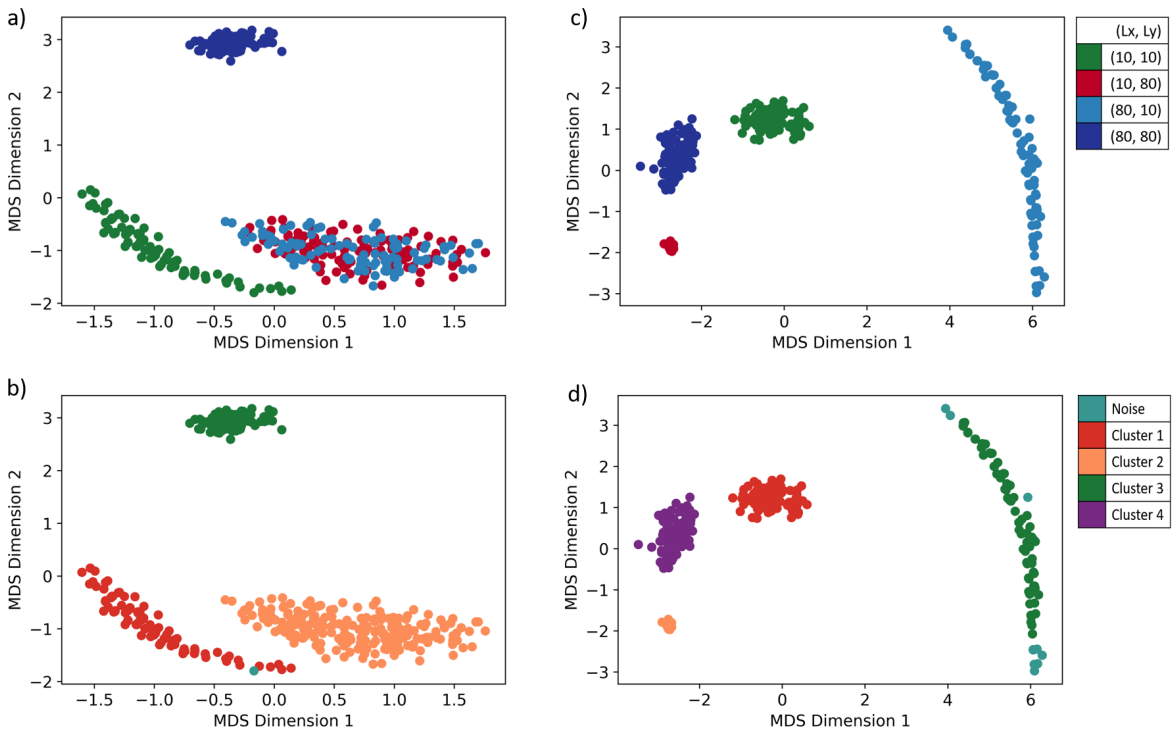


Figure 7: Clustering results for two signature functions on **extreme aspect ratio**: a) low dimensional projection of data using **surface to volume** signature function where points are color-coded using the generation settings, b) the same projection of data as in above panel but color-coded with the indices of clusters (**surface to volume** signature function used to cluster data), c) low dimensional projection using for the signature function of **normalized aspect ratio** where points are color coded with the generation setting, d) the same projection as in panel above but points are color coded with indices of clusters (**aspect ratio** signature function used to cluster data).

and (80,10) are grouped close and separated from the other morphologies with two other settings of grain sizes. The clustering results in panel b) of Figure 7 exhibit a pattern resembling the one described earlier, wherein clusters with morphologies sharing similar grain size and aspect ratio tend to group, effectively distinguishing and separating morphologies with differing grain sizes. Panel c) of Figure 7 shows the separation of data in four different regions when the normalized aspect ratio signature function is used; in this case, the data again is located in four distinct areas of the low dimensional embedding. The clustering results in panel d) of Figure 7 for the normalized aspect ratio signature function cluster the data in four clusters.

To summarize, based on the results for three distinct datasets, COMODO shows the ability to capture different morphological features through a configurable signature function. The accurate and reliable separation and clustering of morphologies showcase that morphologies with different morphological features are treated as distinct entities. The presented results also demonstrate that the COMODO operator incorporates information about both the values of the signature function and the graph structure. In particular, both MDS projections and DBSCAN clustering affirm that (i) the signature function captures the geometric similarity but ignores the directionality of anisotropic grains, and (ii) COMODO captures differences in the graph structure. Comparison with other methods of representing morphologies: descriptor-based morphology representation and statistical function representation are included in the Supplementary Information. The additional analysis demonstrates higher Rand Index values over a wider range of cutoff distances for COMODO compared to two other approaches. The plots and associated discussion are included in the SI section.

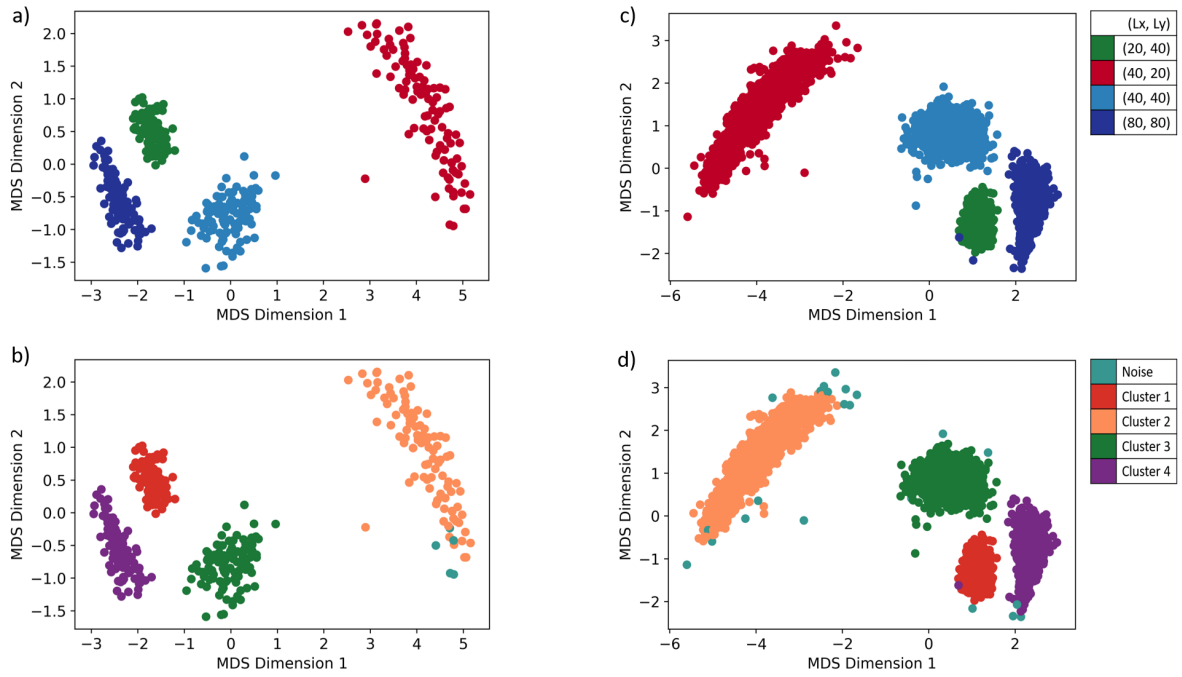


Figure 8: Clustering results for the dataset of varying size and **normalized aspect ratio** signature function: a) low dimensional projection of dataset with **400 morphologies** where points are color-coded using the generation settings, b) the same projection of data as in above panel but color-coded with the indices of clusters, c) low dimensional projection of dataset with **4,000 morphologies** where points are color coded with the generation setting, d) the same projection as in panel above but points are color coded with indices of clusters (**4,000 morphologies**).

3.3 Distance measure is resilience to the size of datasets

To check COMODO sensitivity towards large datasets, the number of morphologies in the datasets is varied. Two datasets with composite morphologies of moderate aspect ratio are used. The first dataset has 400 morphologies with 100 replicas per grain size settings, whereas the second dataset consists of 4,000 morphologies with 1,000 replicas per grain size. All morphologies correspond to the first dataset with a moderate aspect ratio - as listed in the Data Generation subsection. A normalized aspect ratio signature function is used. Panel a) and c) of Figure 8 depict the results in low dimensional projections of datasets with 400 morphologies and 4,000 morphologies, respectively. The data is colored using a data generation scheme, as listed in the legend of the figure. Panel b) and d) depict the clustering results for two datasets with 400 and 4,000 morphologies, respectively. The clustering results consistently separate the morphologies of different aspect ratios irrespective of the number of replicas per grain size configuration. In both cases, only few points are marked as noise points, while remaining points correctly clustered into four groups of points. Moreover, the relative position of clusters is preserved between the two datasets. If we choose the morphologies with grain size $(40, 20)$ (marked red in the top panels) as the reference set, these morphologies are distributed slightly away from the remaining three groups. - in both panels a) and c) Morphologies with grain size $(80, 80)$ are located the furthest away from the reference set, regardless of the size of the dataset. Morphologies for two remaining configurations: $(20, 40)$ and $(40, 40)$ are located in the center yet closer to $(80, 80)$ morphologies, again this positioning is maintained regardless of the dataset size. The larger dataset was the largest analyzed on the desktop machine. The major bottleneck for larger analysis was not the representation or distance calculation but the clustering step.

3.4 Distance measure is resilient to diverse morphologies

In this experiment, COMODO performance on different types of morphologies is evaluated. This is done by preparing a diverse set of morphology data, as explained in the data generation subsection.

Here, five different types of morphologies are used. As a reminder, the grain size configuration, size of the morphology (square and thin film), and types of morphology (spinodal and composite) are varied. Each morphology type has 100 replicas generated, resulting in a total of 500 morphologies included in the analysis. Panel a) of Figure 9 depicts the low dimensional projection of the data with surface-to-volume signature function used to compute distance matrix. Points are color-coded using the true labels set listed in the legend of the figure. Similar to previous results, the true labels correspond to the settings of the morphology generation. Panel b) of Figure 9 shows the low dimensional projection of data with clustering indices also for the surface-to-volume signature function. Similarly, panel c) of Figure 9 represents the low dimensional projection of the data with the min-max aspect ratio signature function used to compute the distance matrix. Points are color-coded using true labels, as shown in the legend. Panel b) of Figure 9 depicts the low dimensional projection with clustering labels for the min-max aspect ratio signature function.

This dataset is the most complex dataset out of all the datasets presented in this paper. Although being more complex than morphologies in the previous experiments, the clustering results show good separation when min-max aspect ratio signature function is used. Five clusters are visually spread apart with only several morphologies marked as noise points (see Figure 9 panel c) and d)). The results from DBSCAN clustering mimic the true labels with the maximum value for the min-max aspect ratio signature function of the Rand Index of 0.90 (see panel b) in Figure S2. As a comparison, when the surface-to-volume signature function is chosen, the maximum Rand Index is 0.75 with a very narrow range of cutoff distance. For this signature function, the challenge to cluster the data is visualized in panel b) of Figure 9. Two types of morphologies belonging to composite data ($(80, 20)$, $(80, 80)$) are clustered together (lilac points belonging to cluster 5). Moreover, composite morphologies with grain size $(10, 80)$ (red points in the panel a), are mostly identified as noise points. When the same dataset is analyzed with min-max aspect signature our distance measure captures sufficient information to cluster the data in line with true labels.

4 Conclusion

In conclusion, COMODO as a distance operator demonstrates significant potential for morphology comparison applications due to its configurability, scalability, and adaptability to handle large and diverse datasets. Its ability to be configured through a signature function allows for customization based on specific requirements. The customization of COMODO helps to tailor the task at hand to the characteristics of the morphology and target application. The two-step process of representing morphology as a graph to capture the relative arrangement of components and then capturing the local information of each component through the signature function reduces the dimensionality of the morphology, addressing the challenge of comparing the high dimensional data points. COMODO is scalable to handle large and diverse datasets. Overall, this work showcased COMODO as a versatile tool for analyzing and understanding the relationships and similarities within complex morphology datasets, supporting a range of applications in the field of material science.

In future work, we plan to use this measure to organize morphologies in databases, where the configurablility of the distance operator can aid in efficient indexing and retrieval of data based on similarity. Moreover, we expect the distance operator to be used in inverse design processes, where the COMODO establishes the similarity between the target morphology and the morphologies already screened.

Data availability

The datasets generated, and the code is available in the repository https://github.com/owodolab/COMODO.

Acknowledgements This work was supported by the National Science Foundation (1906344 and 1910539). All authors acknowledge the support provided by the Center for Computational Research at the University at Buffalo.

References

[1] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

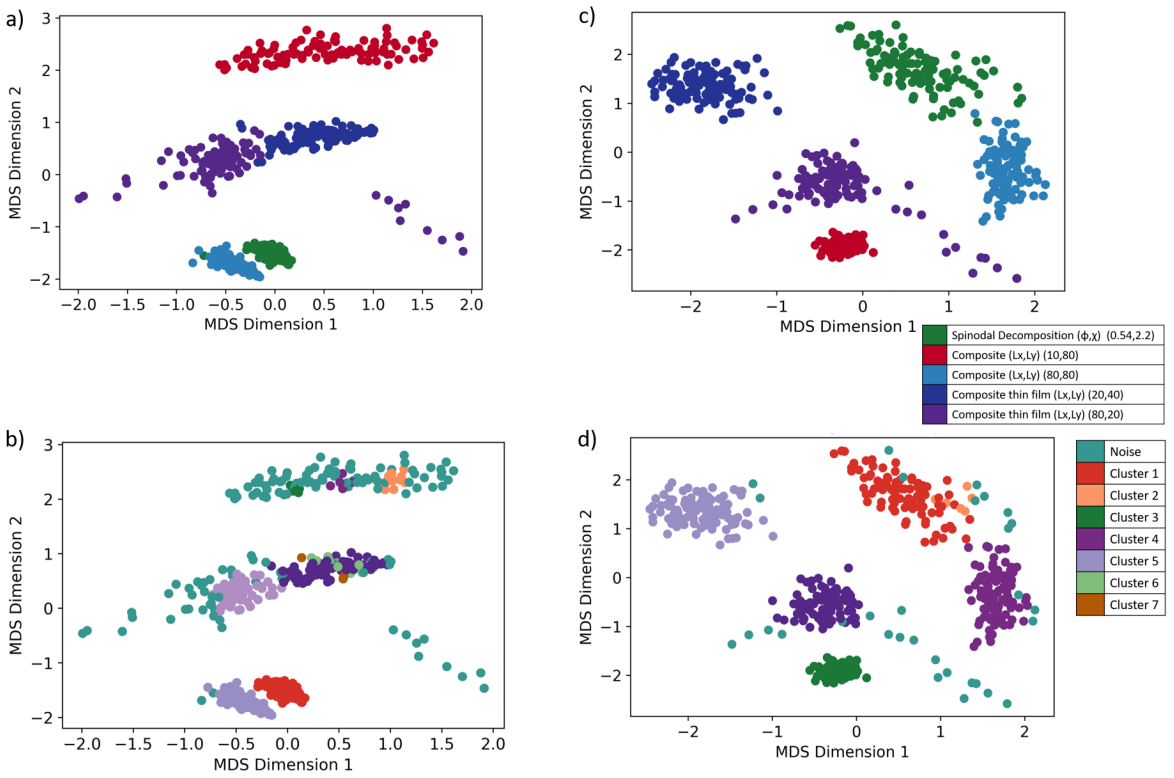


Figure 9: Clustering results of for **diverse morphology dataset**;: a) low dimensional projection of data using **surface to volume** signature function where points are color-coded using the generation settings, b) the same projection of data as in above panel but color-coded with the indices of clusters (**surface to volume** signature function used to cluster data), c) low dimensional projection using for the signature function of **min-max aspect ratio** where points are color coded with the generation setting, d) the same projection as in panel above but points are color coded with indices of clusters (**min-max aspect ratio** signature function used to cluster data).

[2] Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. On the surprising behavior of distance metrics in high dimensional space. In *Database Theory—ICDT 2001: 8th International Conference London, UK, January 4–6, 2001 Proceedings 8*, pages 420–434. Springer, 2001.

[3] Ivano Benedetti and Vincenzo Gulizzi. A grain-scale model for high-cycle fatigue degradation in polycrystalline materials. *International Journal of Fatigue*, 116:90–105, 2018.

[4] Ramin Bostanabad, Yichi Zhang, Xiaolin Li, Tucker Kearney, L Catherine Brinson, Daniel W Apley, Wing Kam Liu, and Wei Chen. Computational microstructure characterization and reconstruction: Review of the state-of-the-art techniques. *Progress in Materials Science*, 95:1–41, 2018.

[5] Victor M Calo, Oleg Iliev, Zahra Lakdawala, KHL Leonard, and Galina Printsypar. Pore-scale modeling and simulation of flow, transport, and adsorptive or osmotic effects in membranes: The influence of membrane microstructure. *International journal of advances in engineering sciences and applied mathematics*, 7:2–13, 2015.

[6] Ahmet Cecen, Tony Fast, and Surya R Kalidindi. Versatile algorithms for the computation of 2-point spatial correlations in quantifying material structure. *Integrating Materials and Manufacturing Innovation*, 5:1–15, 2016.

[7] Ahmet Cecen, Berkay Yucel, and Surya R Kalidindi. A generalized and modular framework for digital generation of composite microstructures. *Journal of Composites Science*, 5(8):211, 2021.

[8] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms*. MIT press, 2022.

[9] Ahed Elmsallati, Connor Clark, and Jugal Kalita. Global alignment of protein-protein interaction networks: A survey. *IEEE/ACM transactions on computational biology and bioinformatics*, 13(4): 689–705, 2015.

[10] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[11] Christophe Fiorio and Jens Gustedt. Two linear time union-find strategies for image processing. *Theoretical Computer Science*, 154(2):165–181, 1996.

[12] Ramiro García-García and R Edwin García. Microstructural effects on the average properties in porous battery electrodes. *Journal of Power Sources*, 309:11–19, 2016.

[13] Martin L Green, Benji Maruyama, and Joshua Schrier. Autonomous (ai-driven) materials science, 2022.

[14] Anthony JG Hey, Stewart Tansley, Kristin Michele Tolle, et al. *The fourth paradigm: data-intensive scientific discovery*, volume 1. Microsoft research Redmond, WA, 2009.

[15] Namit Juneja, Jaroslaw Zola, Varun Chandola, and Olga Wodo. Graph-based strategy for establishing morphology similarity. In *33rd International Conference on Scientific and Statistical Database Management*, pages 169–180, 2021.

[16] JB Kruskal and M Wish. Multidimensional scaling sage publications beverly hills, 1978.

[17] Hao Liu, Berkay Yucel, Daniel Wheeler, Baskar Ganapathysubramanian, Surya R Kalidindi, and Olga Wodo. How important is microstructural feature selection for data-driven structure-property mapping? *MRS Communications*, 12(1):95–103, 2022.

[18] Tim Mueller, Aaron Gilad Kusne, and Rampi Ramprasad. Machine learning in materials science: Recent progress and emerging applications. *Reviews in computational chemistry*, 29:186–273, 2016.

[19] Balaji Sesha Sarath Pokuri, Sambuddha Ghosal, Apurva Kokate, Soumik Sarkar, and Baskar Ganapathysubramanian. Interpretable deep learning for guided microstructure-property explorations in photovoltaics. *npj Computational Materials*, 5(1):1–11, 2019.

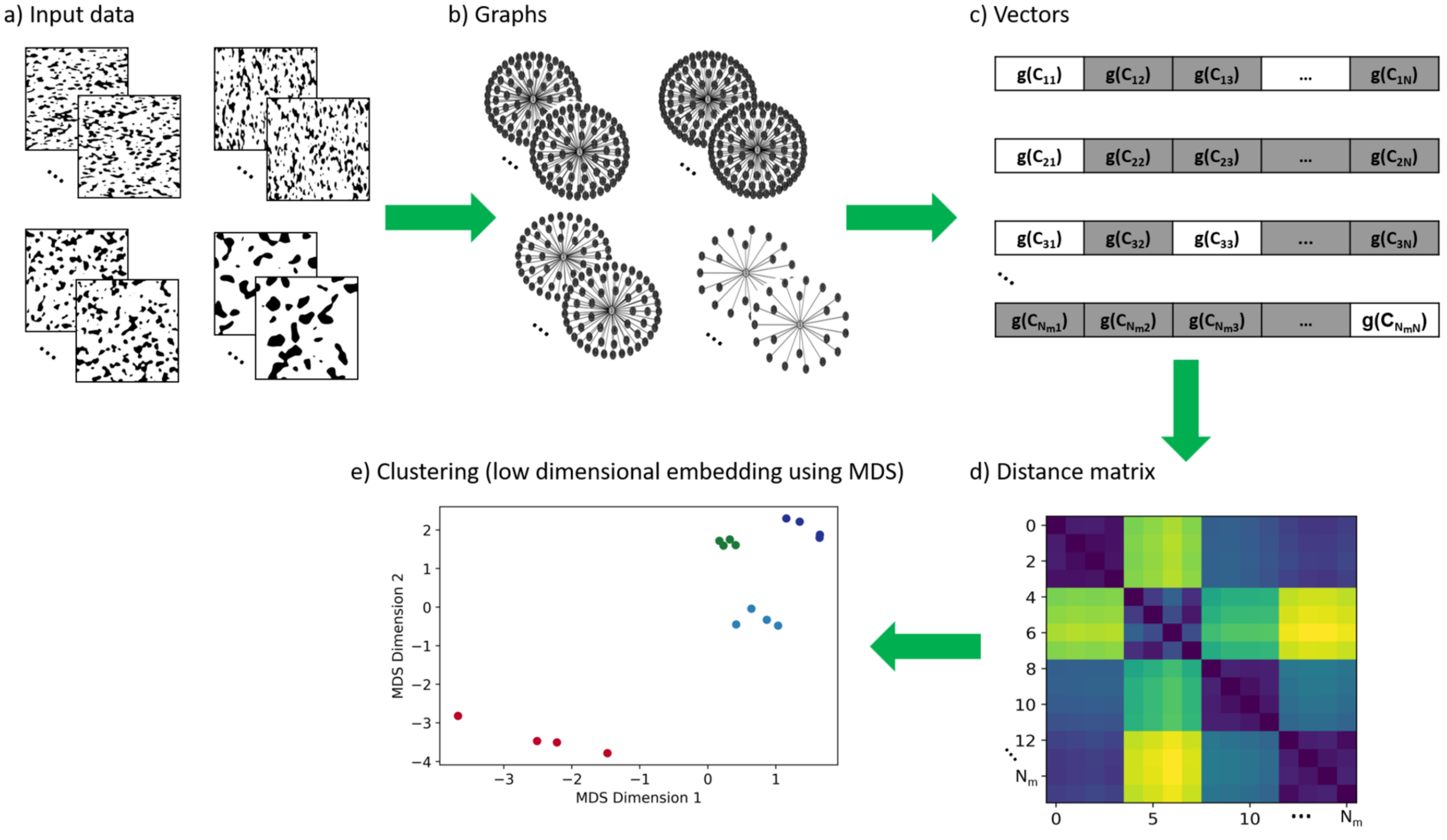
[20] Shuyu Qin, Yichen Guo, Alibek T Kaliyev, and Joshua C Agar. Why it is unfortunate that linear machine learning "works" so well in electromechanical switching of ferroelectric thin films. *Advanced Materials*, 34(47):2202814, 2022.

[21] Olga Wodo and Baskar Ganapathysubramanian. Computationally efficient solution to the cahn–hilliard equation: Adaptive implicit time schemes, mesh sensitivity analysis and the 3d isoperimetric problem. *Journal of Computational Physics*, 230(15):6037–6060, 2011.

[22] Olga Wodo, John D Roehling, Adam J Moulé, and Baskar Ganapathysubramanian. Quantifying organic solar cell morphology: a computational study of three-dimensional maps. *Energy & Environmental Science*, 6(10):3060–3070, 2013.

[23] Kesheng Wu, Ekow Otoo, and Arie Shoshani. Optimizing connected component labeling algorithms. In *Medical Imaging 2005: Image Processing*, volume 5747, pages 1965–1976. SPIE, 2005.

[24] Chih-Hsuan Yang, Balaji Sesha Sarath Pokuri, Xian Yeow Lee, Sangeeth Balakrishnan, Chinmay Hegde, Soumik Sarkar, and Baskar Ganapathysubramanian. Multi-fidelity machine learning models for structure–property mapping of organic electronics. *Computational Materials Science*, 213:111599, 2022.


a) Input data
b) Graphs
c) Vectors
g(C11) g(C12) g(C13) ... g(C1N)
g(C21) g(C22) g(C23) ... g(C2N)
g(C31) g(C32) g(C33) ... g(C3N)
g(CNm1) g(CNm2) g(CNm3) ... g(CNmN)
d) Distance matrix
e) Clustering (low dimensional embedding using MDS)
MDS Dimension 1
MDS Dimension 2