Correlative Analysis of Metal Organic Framework Structures through Manifold Learning of Hirshfeld Surfaces

Xiaozhou Shen, Tianmu Zhang, Scott R. Broderick, and Krishna Rajan

Dept. of Materials Design and Innovation, University at Buffalo - The State University of New York

1 Workflow and Data

1.1 Crystal Structure and Hirshfeld Surface

Hirshfeld surface analysis and surface properties were calculated by the Tonto computational chemistry package (backend of CrystalExplorer[1]), which is based on the structural data of MOF's taken from the Cambridge Crystallographic Data Center (CCDC)[2]. The calculated results contain triangulated positions of the Hirshfeld surface and surface properties including d_{i}, d_{e} and d_{norm}.

1.2 2D Fingerprint Plots

The 2D fingerprint plots are, by definition, the 2D histograms of the d_{i} and d_{e} values for all of the points on the Hirshfeld surface. In the correlative analysis presented in this work, the 2D histograms are treated as 2D images, and thus 2D arrays. To compare a set of 2D fingerprint plots from MOF structures of various atomic/molecular sizes and compositions, all of the histograms use the same range and bin values for both the d_{i} and d_{e} axes. The lower limit of both the d_{i} and d_{e} axes are set to 0 since there will be no d_{i} and d_{e} values below 0. The upper limit is set to be the maximum of all of the d_{i} and d_{e} values from the data set, with a small margin (0.1 Angstrom) added to ensure the largest d_{i} and d_{e} values for a bin with non-zero counts will not be just on the edge of the 2D image. The 2D histograms are further normalized such that the sum of the counts of all of the bins times the bin size is equal to one.

1.3 Mapping of the Fingerprint Plots to Embedding and Graph

The input of the Isomap algorithm can be the coordinates of a set of points in Euclidean space or a distance matrix representing some metric defined on the set of points. As the 2D fingerprint plots are treated as 2D arrays, the Euclidean distances between each pair from the set of plots is used in this study. The direct output of Isomap is a low-dimensional representation of the input points, i.e., each original data point becomes a point in a low dimensional Euclidean space. With information of the k-nearest neighbors for each point, an unweighted undirected graph can be generated, and this graph is the basis for the correlative material search and design.

2 MOF Descriptors

In this section we examine the correlation between the different descriptors for the MOF structures used in this study. The goal is to show that the target properties (surface area and density) are not strongly correlated with the crystallographic properties and to therefore support the argument that the correlative analysis performed in this work reveals information that is otherwise not easily discovered by other methods. In Figure S1, we show the heat map of the correlation matrix of the descriptors. The pore size/surface area related properties are relatively highly correlated, but the correlations with the crystallographic descriptors are relatively low.

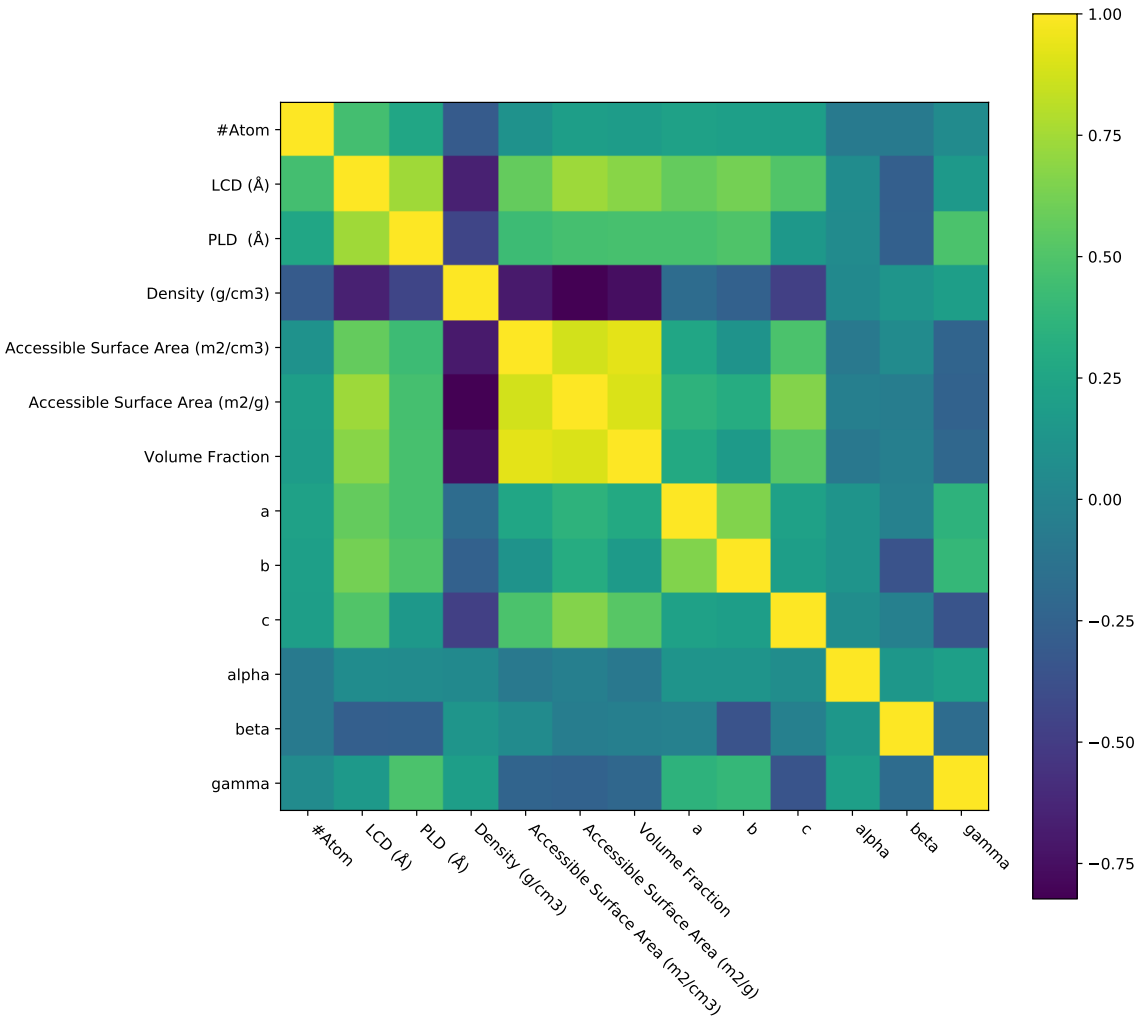

Figure S1: Heat map of the correlation matrix of selected descriptors.

3 Isomap

The implementation of the Isomap algorithm is summarized below (please refer to [3] for details). There are three main steps involved in obtaining the low dimensional projection of the points from their original input space.

- Construct a graph from the given set of points with a defined metric.
- Determine the shortest graph distance on the graph for each pair of points in the input set.
- Construct the low-dimensional representation of the points set based on the graph distances between points.

Given a set of input points, which are assumed to reside on some unknown d-manifold that is embedded in an ambient space with sufficient dimensions, the first step is to calculate the pair-wise distance of the point set in the ambient space. Using the calculated distances, a graph is generated with the vertices being the input data points and with the restriction that each point is allowed to only have edges between itself and its k nearest neighbors. The selection of a proper number of k is crucial for correctly representing the unknown manifold. The graph can be weighted and the weights of the edges are the distances between the points in the ambient space. Shortest path search algorithms such as Floyd-Warshall or Dijkstras can be applied to this graph to obtain the estimated pair-wise geodesic distances for all of the pairs of points. In the last step, multi-dimensional scaling is applied to the estimated geodesic distance matrix to obtain the low-dimensional projection of the input data points.

3.1 Reconstruction Error

The last step of Isomap is to reconstruct a low dimensional representation of the data by using the estimated geodesic distance matrix, and the loss of this reconstruction can be quantified by the Residual Variance, (see note 42 of [3]), i.e., the portion

of the variance in the estimated geodesic distances that is not "explained" by the variance in the low-dimensional projected data.

$$Loss = 1 - R^2(D_{\text{geo}}, D_{\text{low-dim}})]. \tag{1}$$

The D in Equation (1) is the distance matrix. We used this loss here as a measure to evaluate the selection of the Isomap algorithm input parameters.

3.2 Number of Nearest Neighbors

In general, the selection of the number of nearest neighbors can impact the results of the Isomap analysis in two ways[3, 4, 5]: if the number of nearest neighbors is chosen too small, the number of edges in the generated graph will also be small and the graph can potentially become unconnected; or if the number of nearest neighbors is chosen too large, some edges can "shortcuit" the manifold and cause the pair-wise geodesic distance to be incorrectly measured.

Here we examine the effect of the number of neighbors by plotting the reconstruction residual variance versus the number of neighbors. This is shown in Figure S2, where the number of neighbors of 2 is chosen for the result presented in the main text. As a reference, the graphs with different numbers of nearest neighbors are shown in Figure S3. Overall, the relative

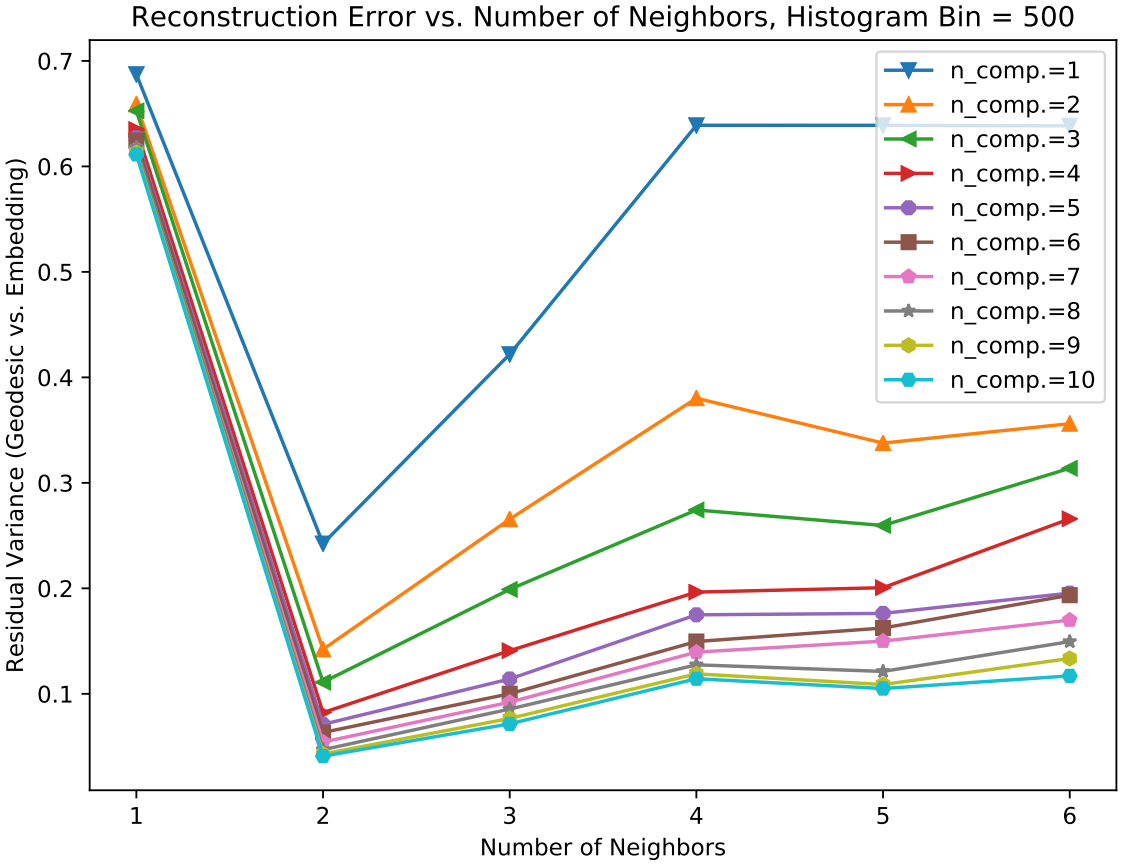

Figure S2: Reconstruction residual variance versus the number of nearest neighbors for different reconstruction dimensions. The number of histogram bins is kept at 500 for all of the curves. For all of the different numbers of reconstruction dimensions studied in this case, the number of nearest neighbors = 2 gives the lowest residual variance. Also, the residual variance decreases as the number of reconstruction dimensions increase.

positions of the points in the graphs with two and three nearest neighbors are similar, and the overall shape of the two graphs are also similar.

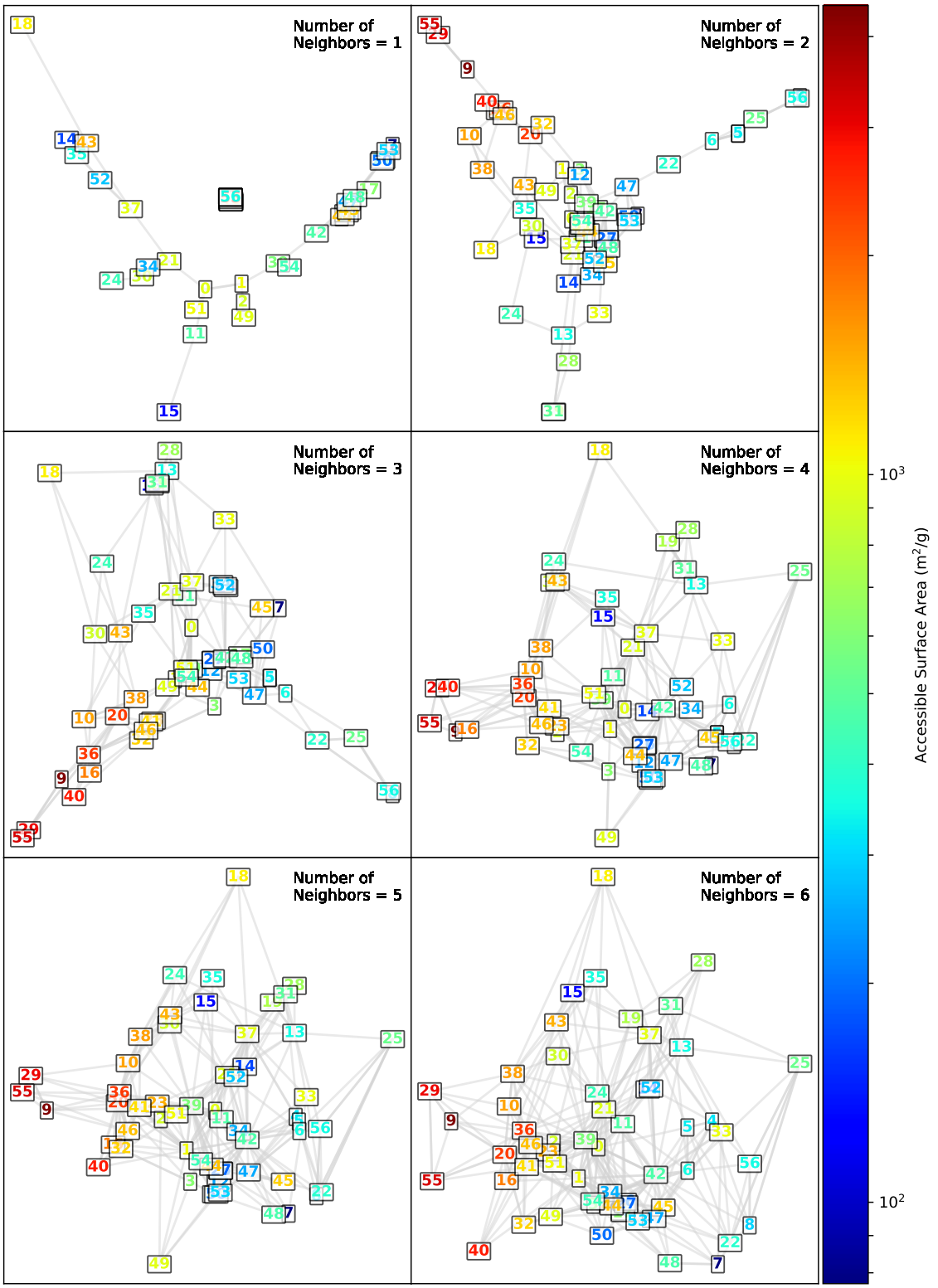

Figure S3: Low-dimensional embedding graph with different numbers of nearest neighbors. The color code of the numbers (MOF structure indices) in every node is for the Accessible Surface Area.

3.3 Number of Dimensions for Low-dimensional Projection

The number of dimensions for the low-dimensional projection can be selected by examining the plot of residual variance vs. the number of reconstruction components. The effect of different numbers of reconstruction dimensions on the residual variance is shown in Figure S4. It is reasonable to choose 2 as the number of reconstruction dimensions.

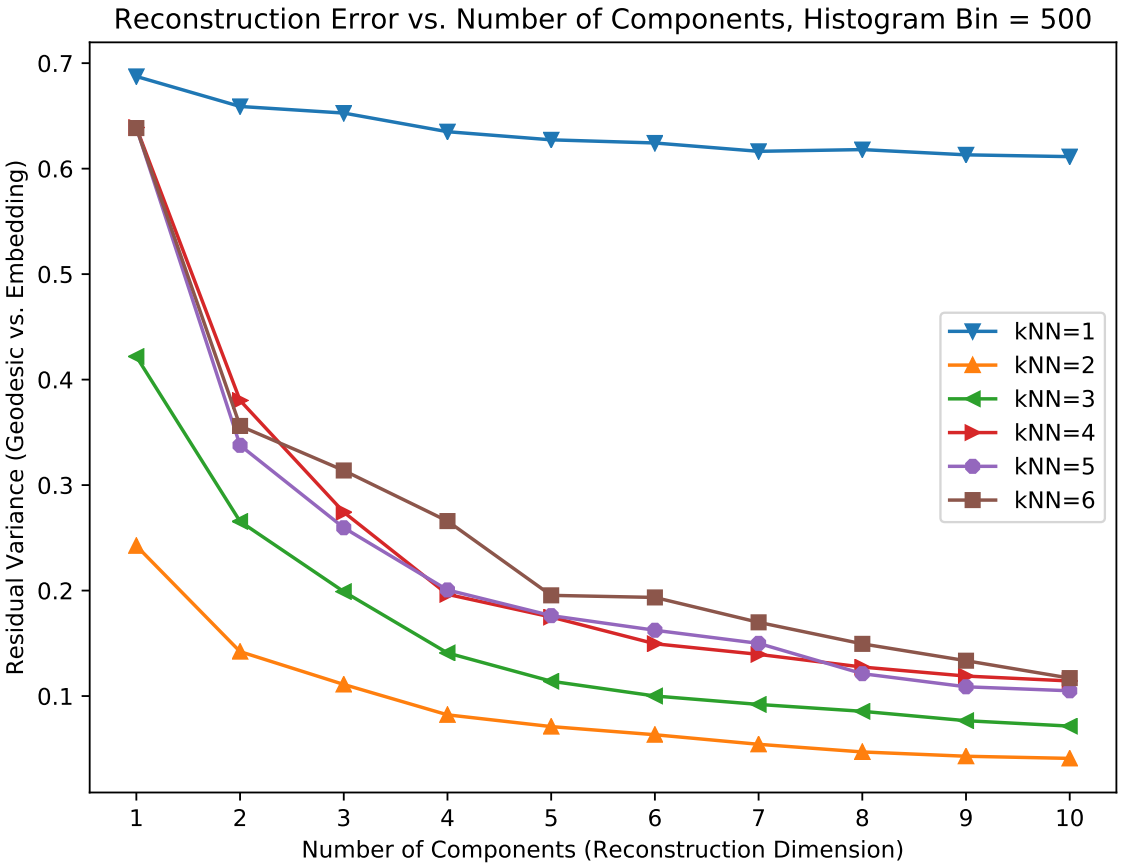

Figure S4: Reconstruction residual variance versus the number of reconstruction dimensions for different numbers of nearest neighbors. The number of histogram bins is kept at 500 for all of the curves. For all of the different number of reconstruction dimensions studied in this case, the number of nearest neighbors = 2 gives the smallest residual variance.

4 Additional Results

4.1 Isomap Graph Network of Input MOF Structures

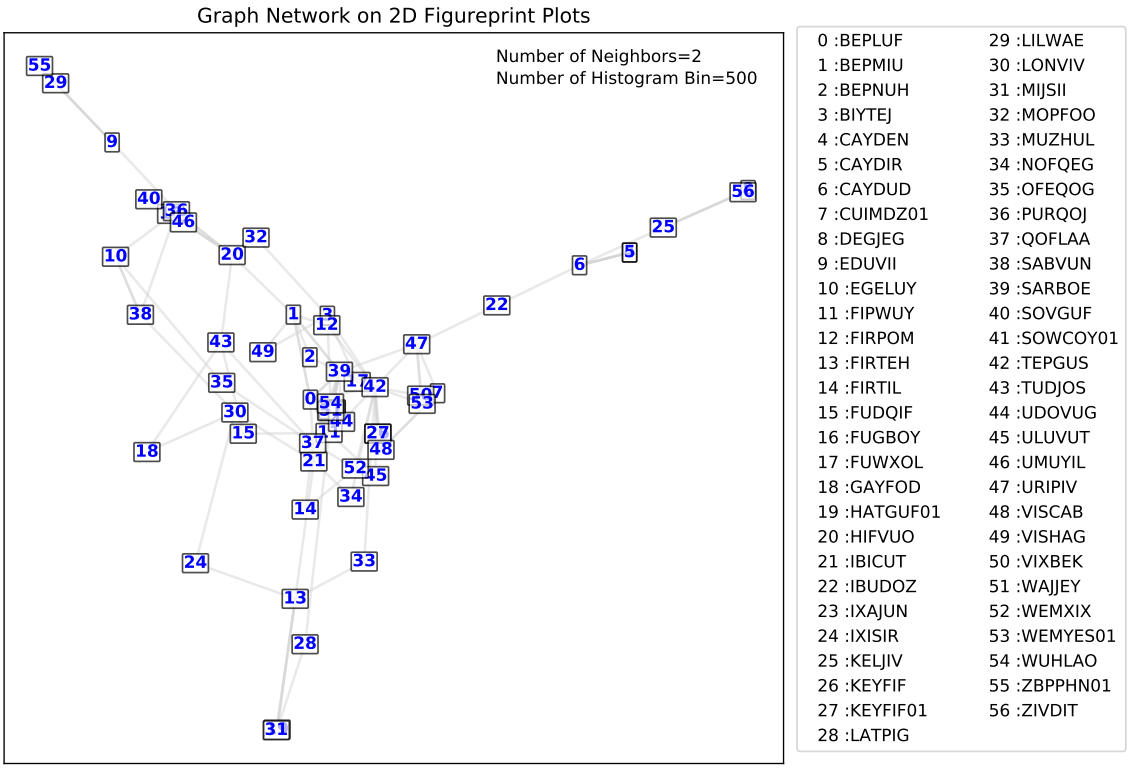

Figure S5: Isomap output: low-dimension embedding and graph network.

4.2 Subgraphs for Different Branches

This section presents the magnified subgraphs on the ends of the three branches in the graph network shown in the main text. The purpose of the magnified subgraphs are to show that for the vertices on each of the ends of the three branches that the edges connect them and therefore potential similarities can be found between them. This conclusion cannot be made solely from the closeness of the vertices in the low-dimensional projection.

4.2.1 Low-dimensional MOF Structures

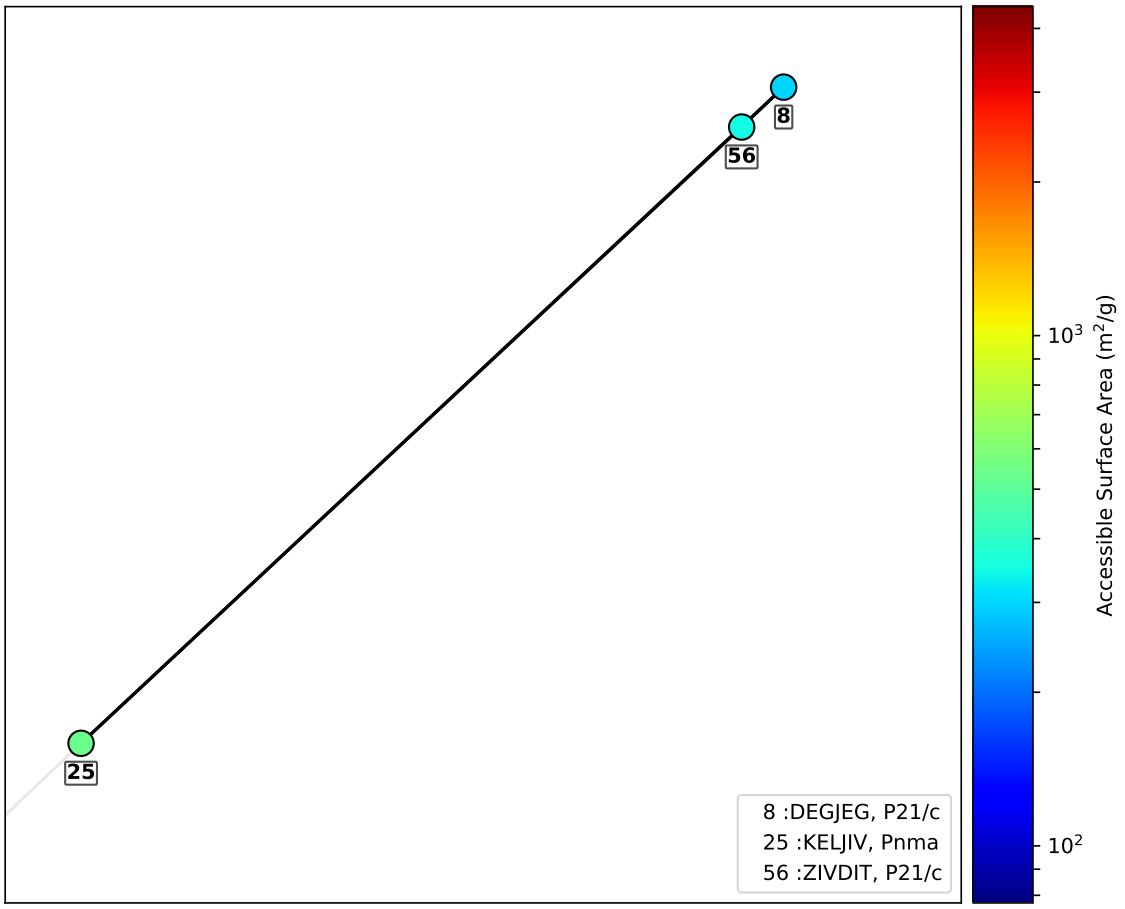

Figure S6: Magnified subgraph showing the vertices and their connecting edges for the three 2D MOF structures.

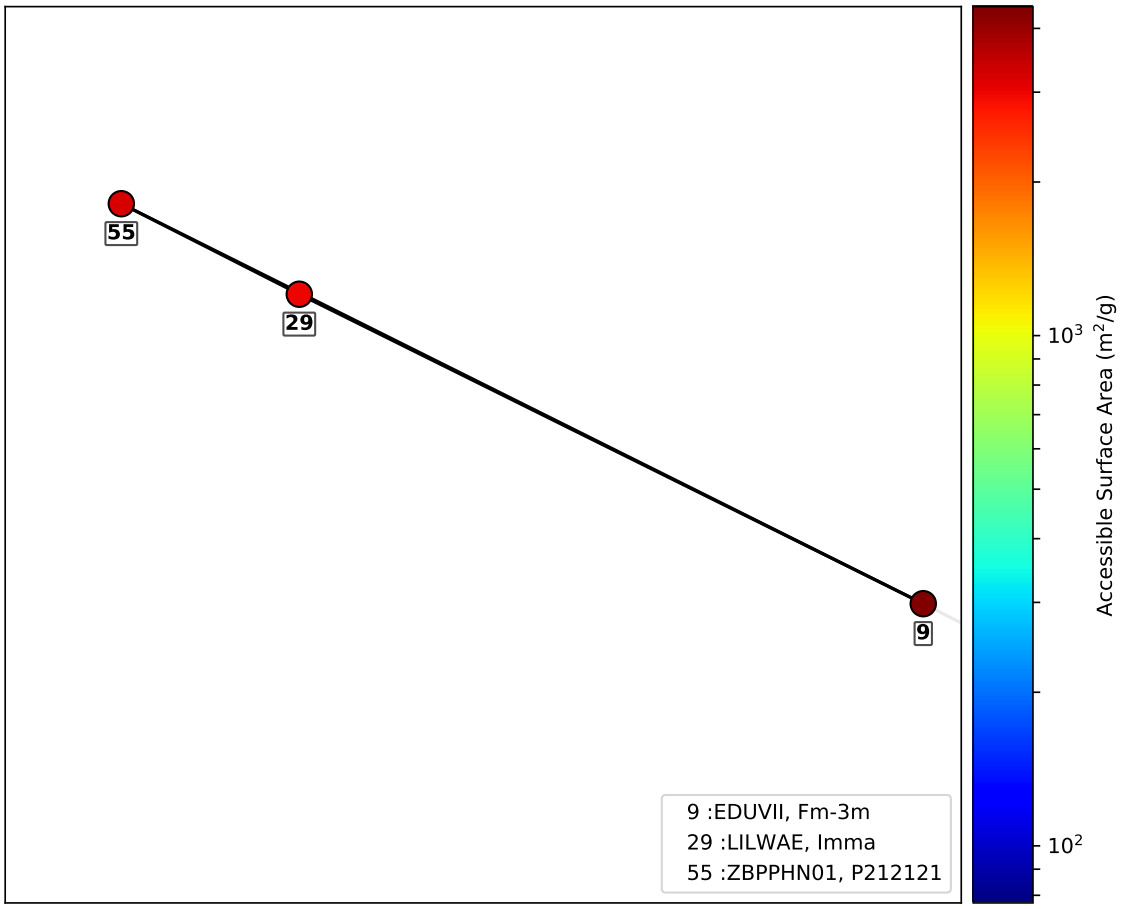

Figure S7: Magnified subgraph showing the vertices and their connecting edges for the two 1D MOF structures.

4.2.2 Cu Coordination

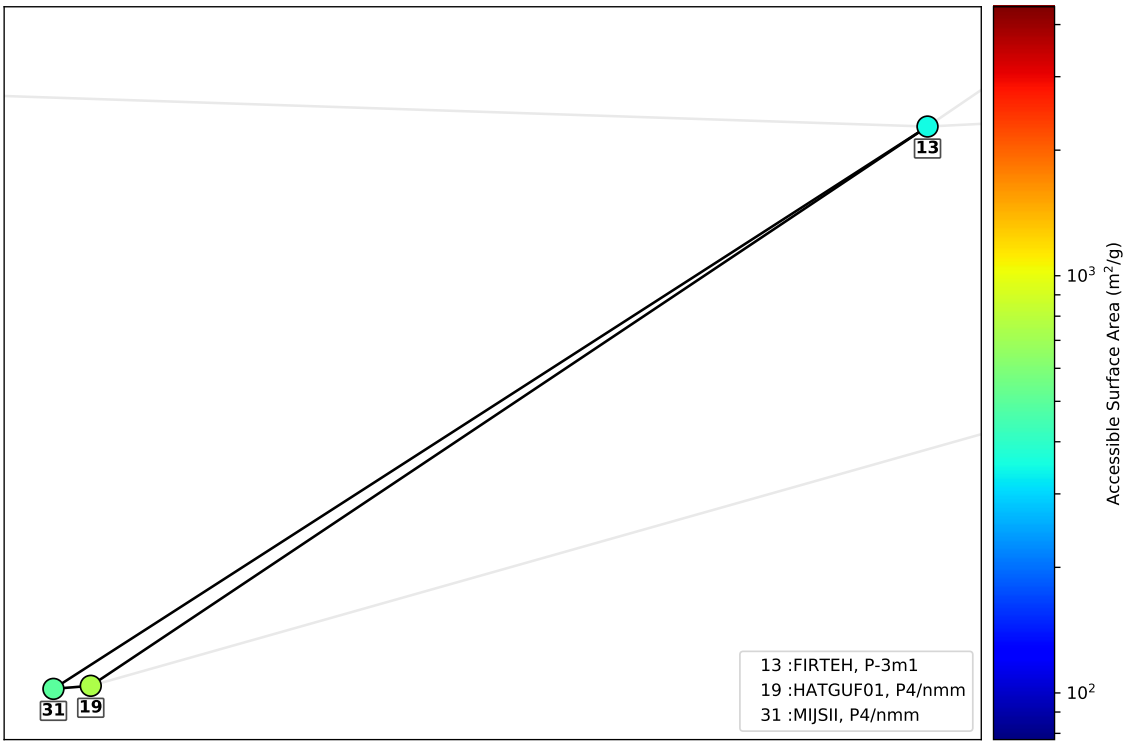

Figure S8: Magnified subgraph showing that the vertices and their connecting edges for the three MOF structures have the same Cu coordination unit.

4.3 Accessible Surface Area

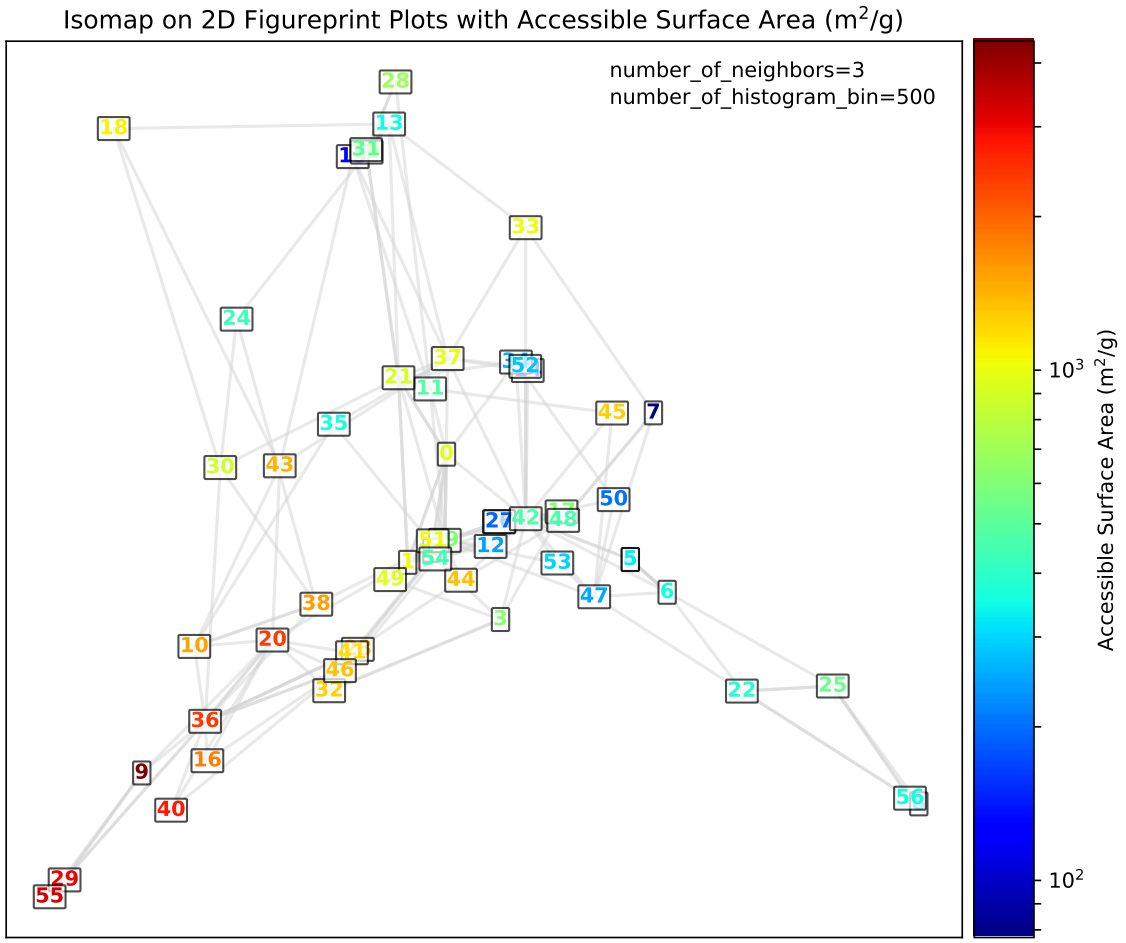

Figure S9: Isomap low-dimension embedding with Accessible Surface Area for a graph constructed with the number of nearest neighbors = 3.

4.4 Density

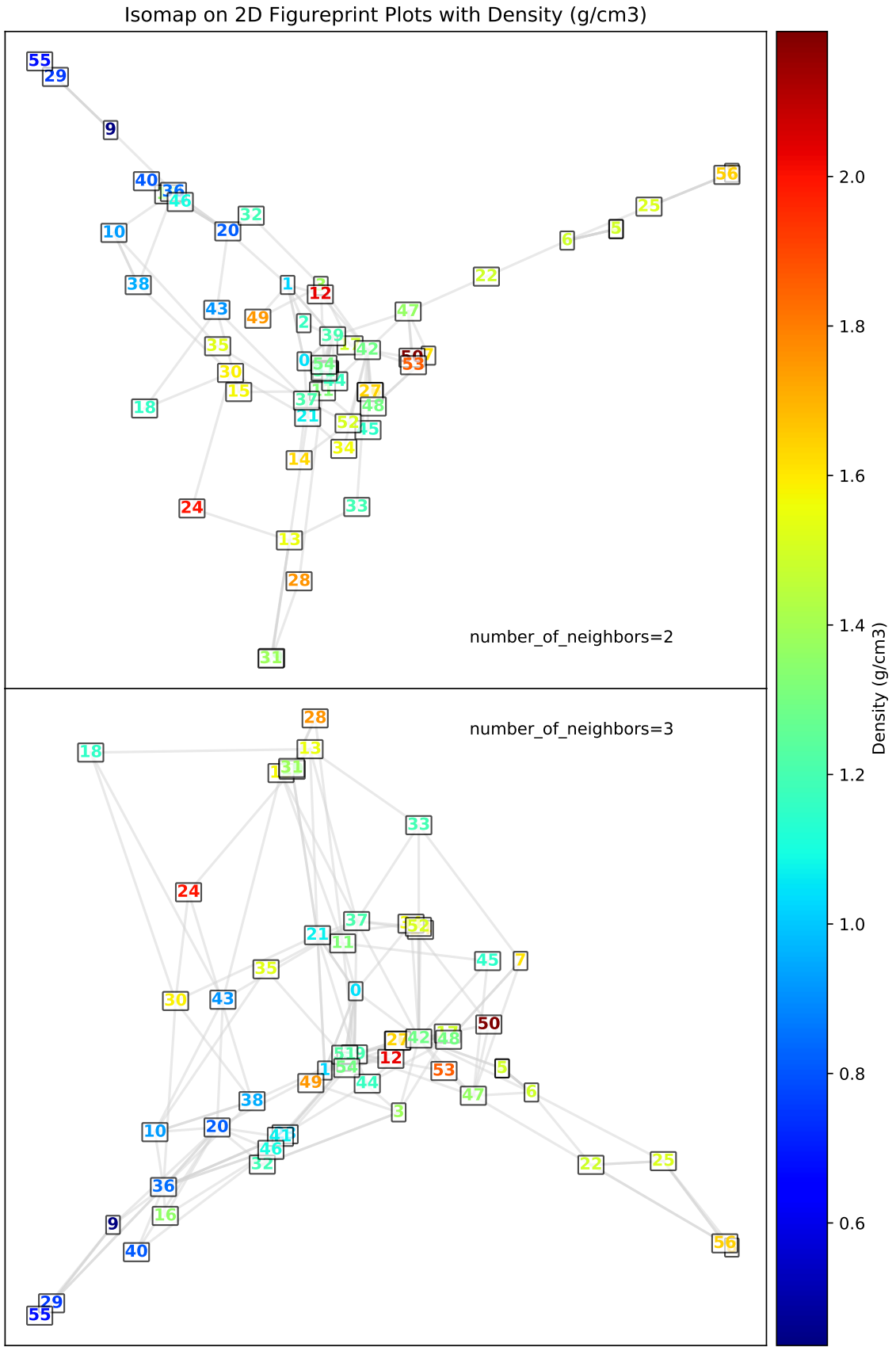

Figure S10: Isomap low-dimension embedding with Density for a graph constructed with the number of nearest neighbors = 2 and 3.

4.5 Volume Fraction

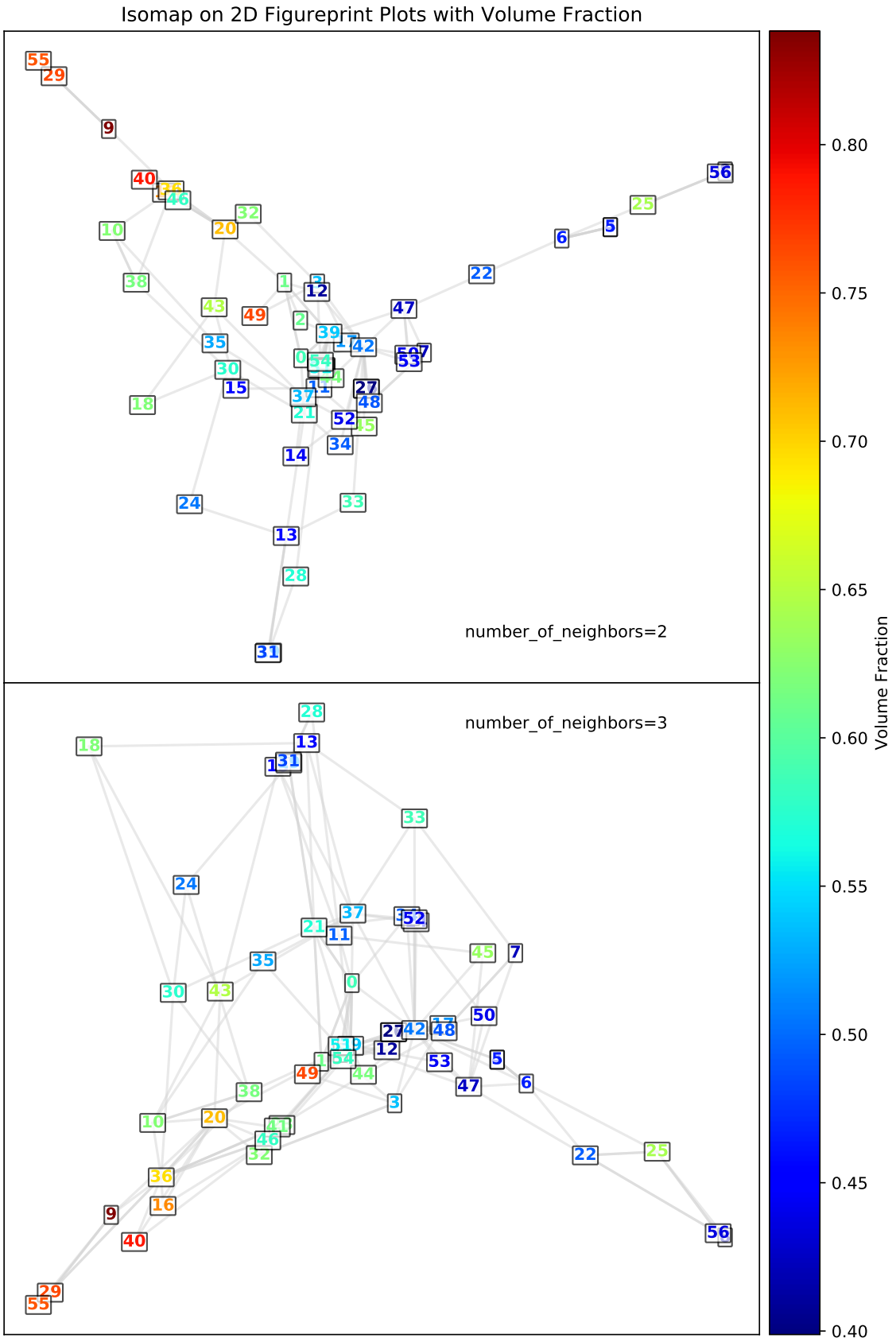

Figure S11: Isomap low-dimension embedding with Volume Fraction for a graph constructed with the number of nearest neighbors = 2 and 3.

5 Effect of 2D Histograms (d_i vs. d_e signature plots) Bin Numbers

In this section, we analyze the effect that the bin numbers of the 2D histogram has on the final results of the Isomap. The 2D histograms (fingerprint plots) are used as the signatures to represent the MOF structures, and the Euclidean distances

between two of the histograms are used by the Isomap algorithm as the metric in the ambient (embedding) space. Different choices of bin number can lead to different Euclidean distances between the signatures for the same pair of MOF structures and thus different final results. In this study for all of the 2D histograms, the value ranges of the d_i and d_e axes are the same and kept fixed (see section 1.2), and the bin edges along the d_i axis are also the same as the bin edges along the d_e axis. The final selection of the number of bins for the 2D histograms is based on the distribution of the pairwise Euclidean distances of all of the input 2D histograms, the residual variance of the Isomap output and the shape of the graph network produced by the Isomap algorithm. That is, we select the number of bins which is in the range where the shape of the final graph network is stable, and the residual variance and the Δ value (see below) are relatively low. The results presented in the main text are obtained by setting the bin number to 500 for both d_i and d_e,

5.1 Distribution of the Distances between 2D Fingerprint Plots

A coarse measure of the distribution of the pair-wise Euclidean distances between all the points in the dataset can be estimated by using[5]

$$\Delta(\vec{x}_i) = \frac{1}{N}\sum_i \frac{\min_{i\neq j}|\vec{x_i} - \vec{x_j}|}{\max_{i\neq j}|\vec{x_i} - \vec{x_j}|}. \tag{2}$$

This value is between 0 and 1, with a smaller value indicating a preferred distance distribution. Figure S12 shows the distribution of the Euclidean distance in the ambient (embedding) space between the pairs of the fingerprint plots for the dataset as well as the corresponding Δ values for different numbers of histogram bins. As the number of bins increases, the overall shape of the distribution does not change much. Nonetheless, the broadness of the distributions and the value of Δ increases as the number of bins increases. This can be explained becasuse with a larger number of bins, the dimension of the space that the histograms are in is also higher. Further, as the dimension of the input space increases, the distance between the farthest and nearest neighbor for a point will decrease[5].

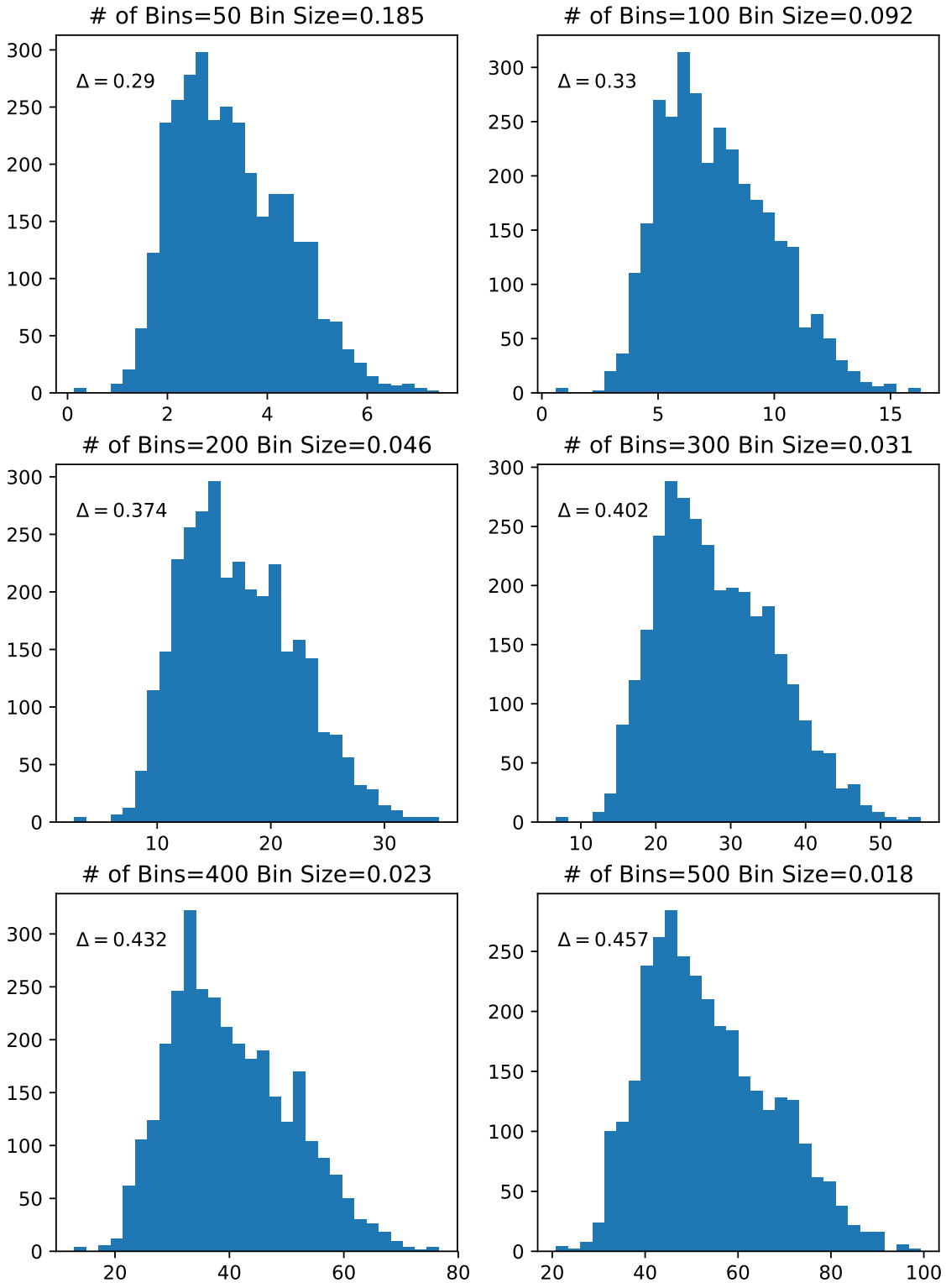

Figure S12: Distributions of the pair-wise distances for all the pairs of the 2D fingerprint plots. The Δ value is a measure of whether a certain distance distribution is preferred.

5.2 Effect of Histogram Bin Size on the Shape of the Graph Network

Figure S13 shows the Isomap produced low-dimensional embedding (graph network) with different numbers of histogram bins. The result is relatively stable as the number of bins increases to 400 and above in terms of the overall shape of the graph network and the relative location of the vertices in the graph. In particular, for the bin number values of 400, 500 and 600, the resulting graph networks all have a 3-branched shape; furthermore, the end vertices of the branches are the same.

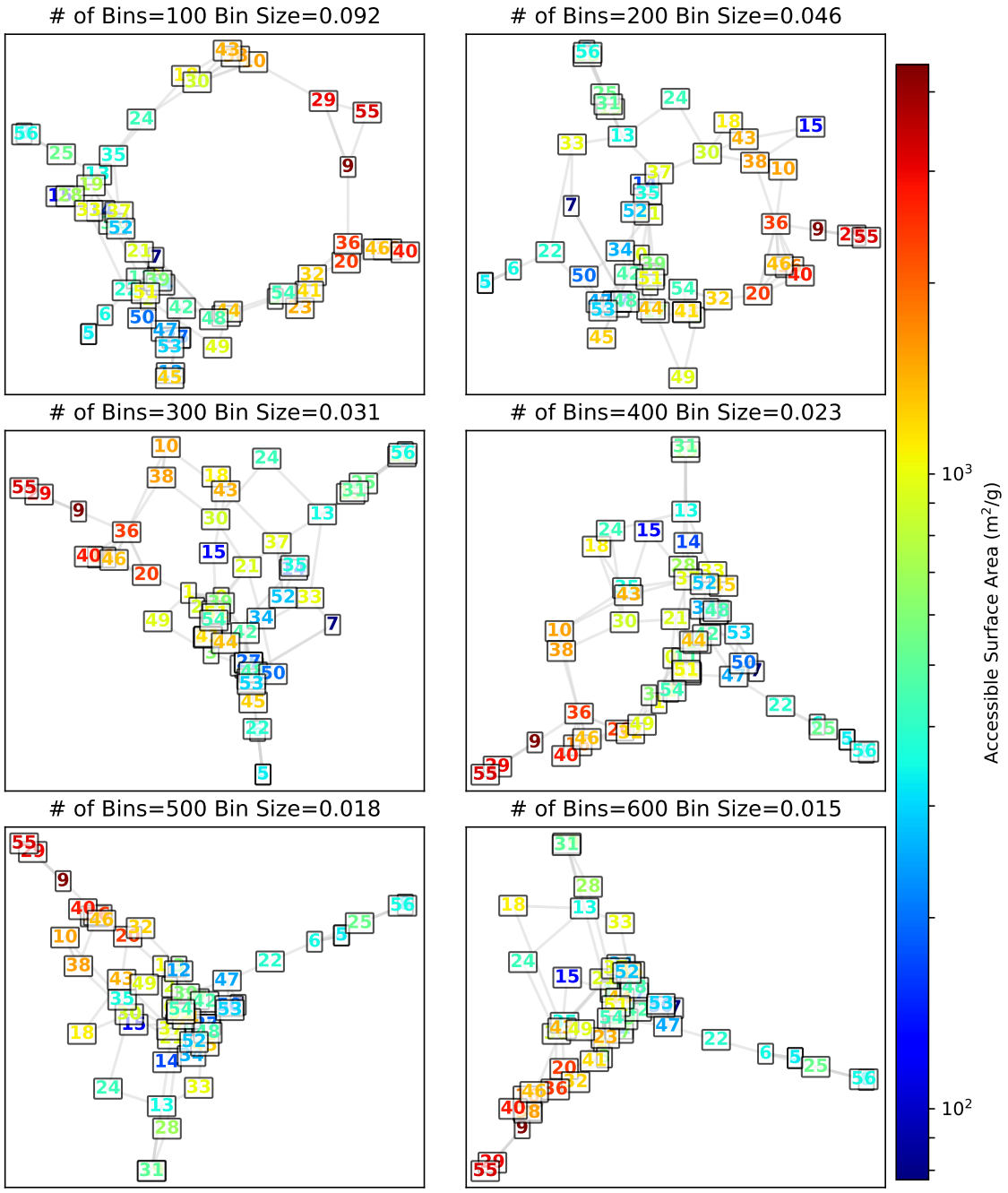

Figure S13: ISOMAP low-dimensional embeddings for different numbers of histogram bins.

5.3 Effect of Histogram Bin Size on Reconstruction Residual Variance

Figure S14 shows the reconstruction residual variance for different number of histogram bins. There is a general decreasing trend of the residual variance as the number of bins increses.

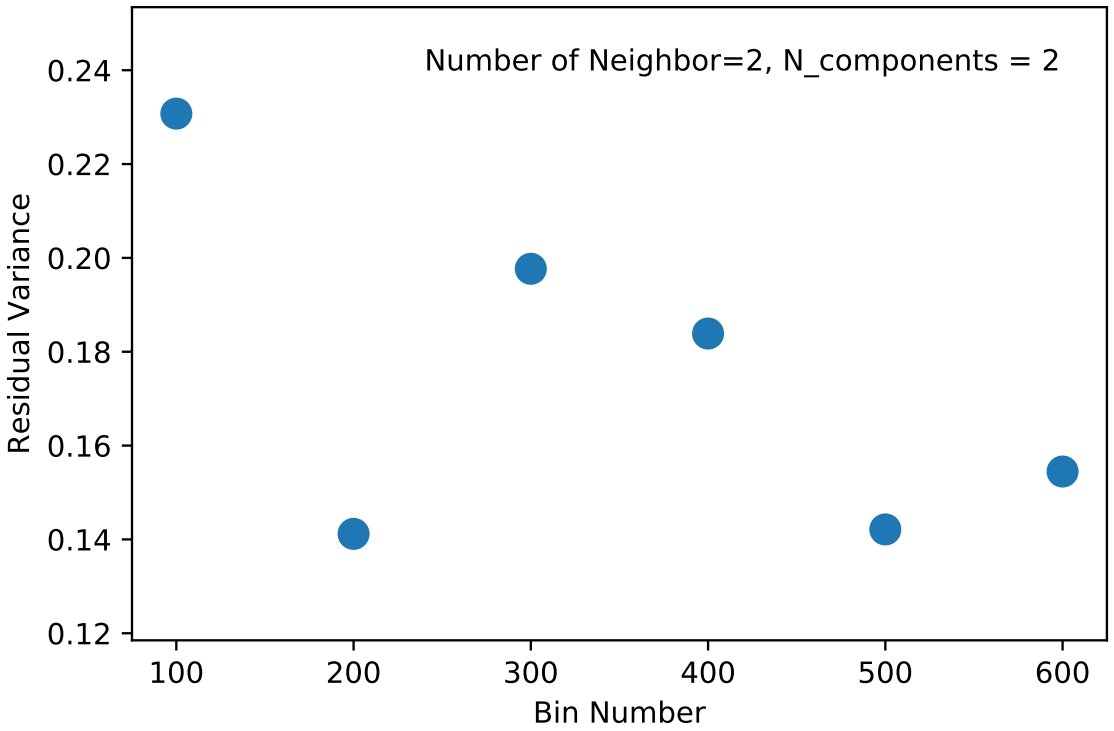

Figure S14: Reconstruction residual variance versus different numbers of histogram bins. The number of nearest neighbors is equal to 2 and the reconstruction dimension is 2.

6 Scalability and Stability

To test the general applicability of this method, we created a larger MOF dataset which has in total 508 different MOF structures with the 57 structures used in the main text included. The purpose is to show that this method is capable of dealing with larger datasets of MOF structures and the results obtained from a smaller sub dataset are relatively stable when the input dataset becomes large.

The procedure for processing the data follows the same as outlined in the previous sections of this Supplementary Materials, with the exception of the cut-off value (upper limit) of the 2D histogram bins, which for all the calculations performed in this section is set to 10 Angstroms.

All of the 508 MOF structures are listed separately in a spreadsheet which is available for download. The spreadsheet includes the CCDC identifier, chemical names and descriptors of the MOF structures.

6.1 Isomap Result of 508 MOF structures

In this section, we present the result of our method applied to all of the 508 MOF structures. The selection and justification for the parameters are based on the discussion in Section 3. The number of nearest neighbors of the Isomap algorithm is set to 6, as it gives the lowest reconstruction residual variance for the reconstruction dimension of 3. In practice, users need to be aware of the fact that a larger number of nearest neighbors will result in more edges being added to the graph network and thus a more complex network. For visualization purposes, the reconstruction dimension is set to 3. In addition, for the sake of clarity, no edge is added to the graph network represented in 3D space. Movies of the 360-degree view of the network at three different altitudes are recorded (available in the online version).

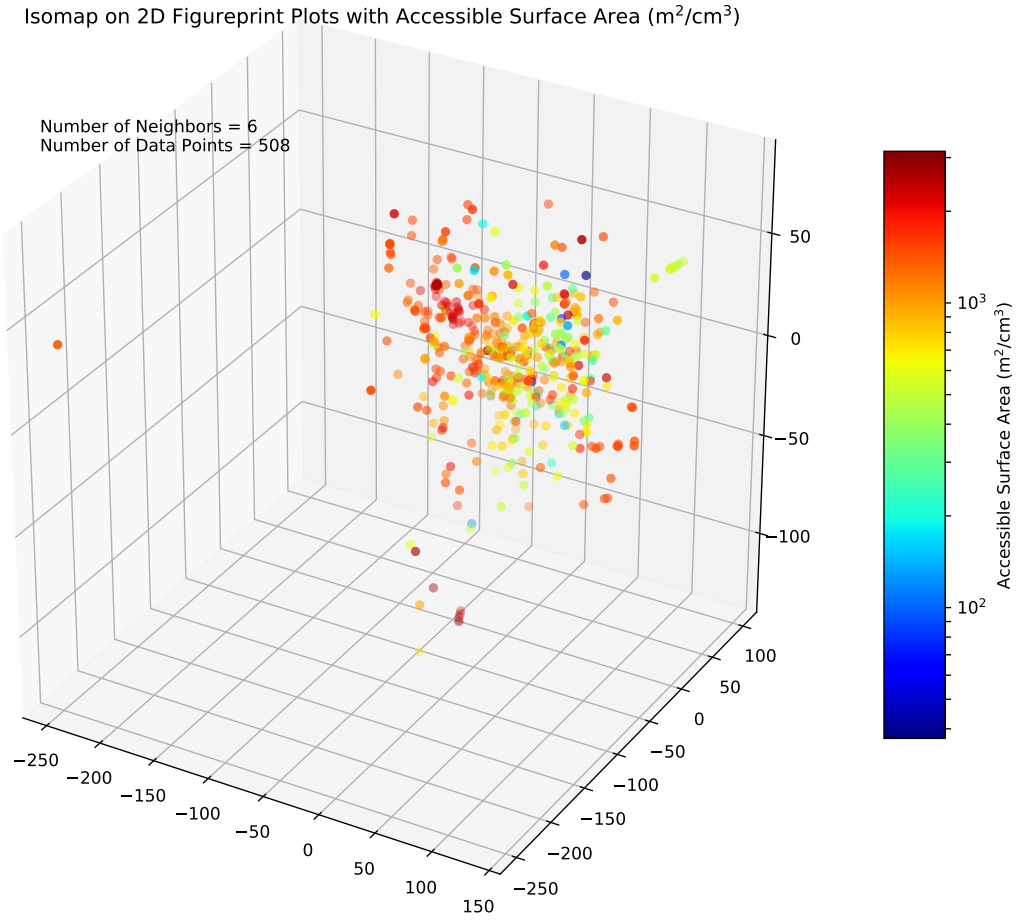

Figure S15: Network of 508 MOF structures show in 3D, without edges. The color bar represents the Accessible Surface Area (m^2/cm^3).

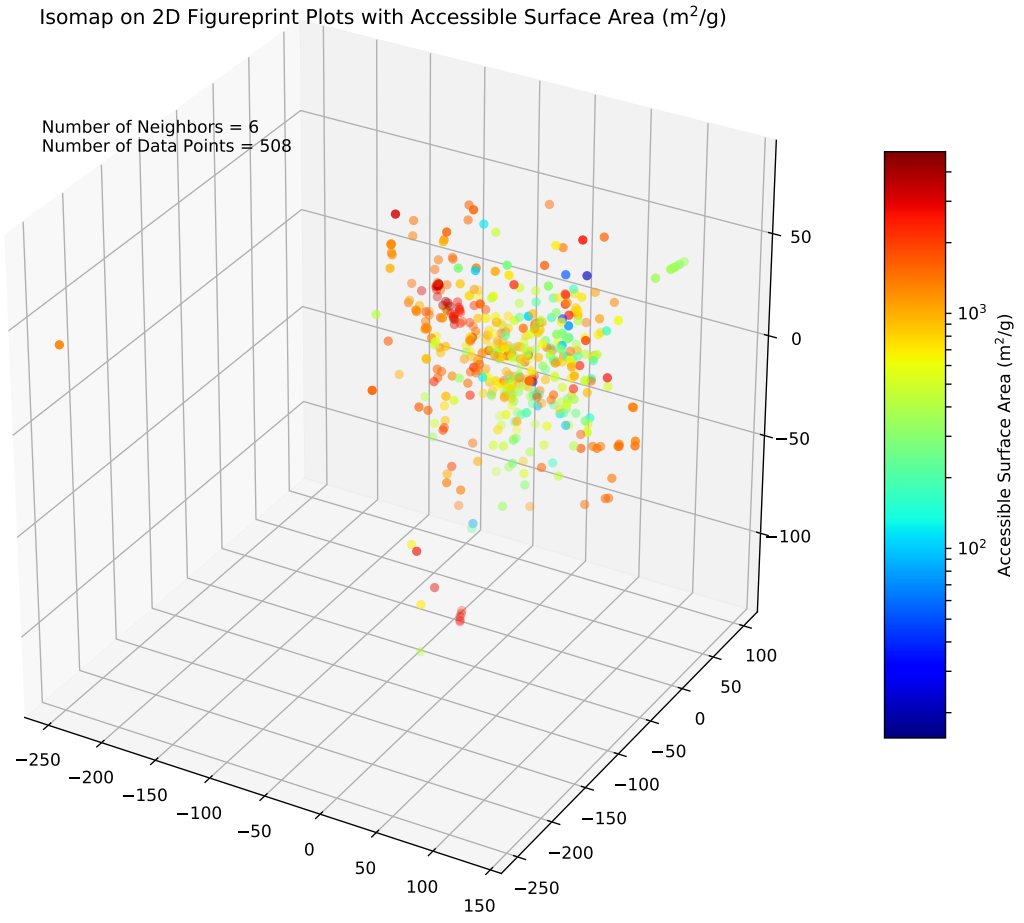

Figure S16: Network of 508 MOF structures show in 3D, without edges. The color bar represents the Accessible Surface Area (m^2/g).

6.2 Scalability and Stability of the Method

In this section we show that the results in the main text, namely the 57 MOF structures, are preserved if the input dataset becomes larger. In the main text, we concluded that our method has identified:

- a global trend on the Accessible Surface Area;
- similar crystal structures (1/2D materials and same coordination geometry);
- a lead compound for the given input dataset.

In the following, we discuss the first two conclusions, while the last one on the lead compound will not be discussed because a lead compound is particular to a given input dataset, and for a different input dataset the lead compound will likely to be different.

6.2.1 Trend on Accessible Surface Area

As the static 3D plots in Section 6.1 cannot fully show the results, we refer the readers to the movie version of the results for a clear representation.

As can be seen from Figure S15 and S16 and their corresponding movies, the 3D network can be divided into two parts: the inner part and the outer part which contains points that surround the inner part. The trend of Accessible Surface Area can be seen on the inner part, i.e. on one side the points are colored in the red region and on the other side the points are colored with yellow and green colors. Also, the points on the high Accessible Surface Area side are closer to each other than the points on the low Accessible Surface Area side. Around the inner part there are scattered points that do not quite follow the trend on the Accessible Surface Area. Future studies will be performed on these individual MOF structures.

6.2.2 Similar Crystal Structures

The case studies here are for the same Cu coordinate geometry and the same 1/2D materials.

We chose randomly, from the dataset of 508 MOF structures 150, 250, 350 and finally all of the 508 MOF structures. In each case, the 57 MOF structures studied in the main text are always included, and in this way their relative relation in the new dataset can be examined and compared with the relative relations in the original dataset. The random numbers of chosen indices of the MOF structures in the dataset are available in the format of Excel in the online version. Figure S17 to Figure S20 show the results of the different datasets. From those figures we can see that for the different case studies in the main text, the compounds in the same case are always close to each other in the new network. This means the similarities among these MOF structures will also be able to be identified if one would start with a different input dataset containing them. Also, the general trend on the Accessible Surface Area is shown in each of the networks. The reconstruction dimension is set to 3 for all of the networks, and the number of nearest neighbors is chosen so the residual variance is the lowest as per Section 6.3.

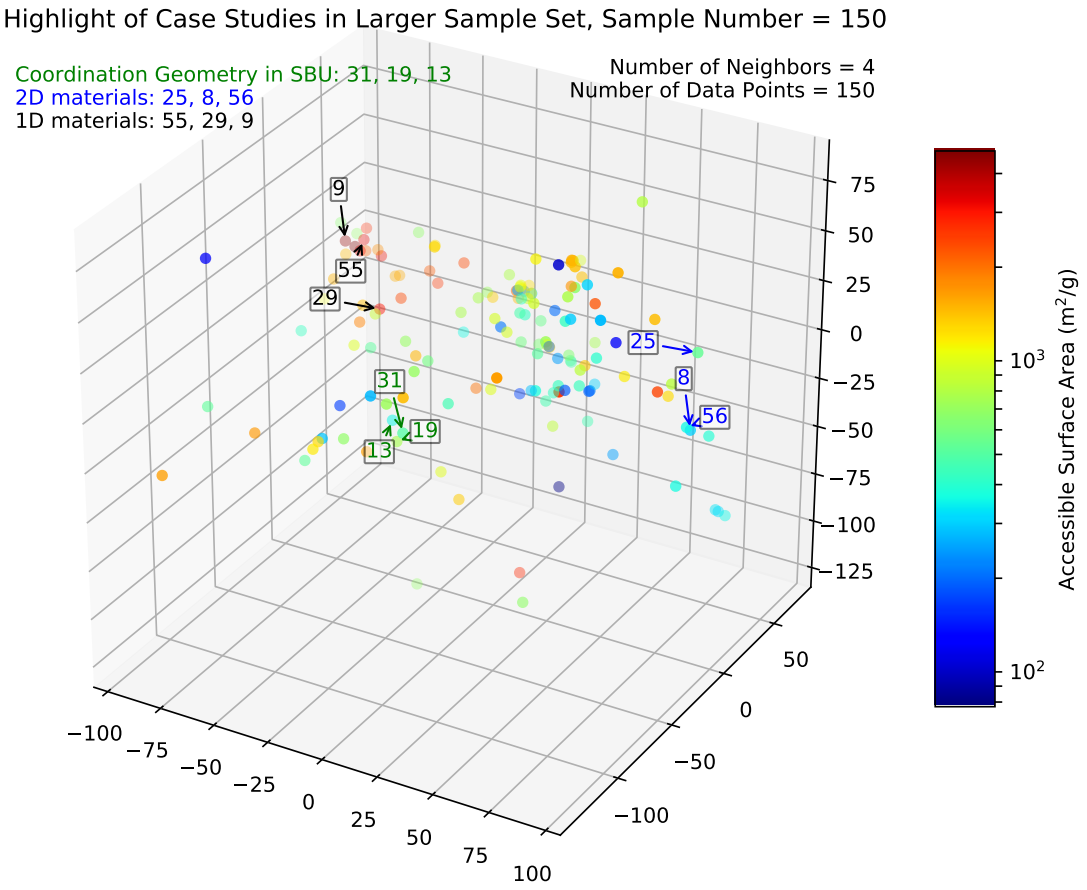

Figure S17: Network of 150 MOF structures show in 3D, without edges. The MOF structures in the case studies are annotated.

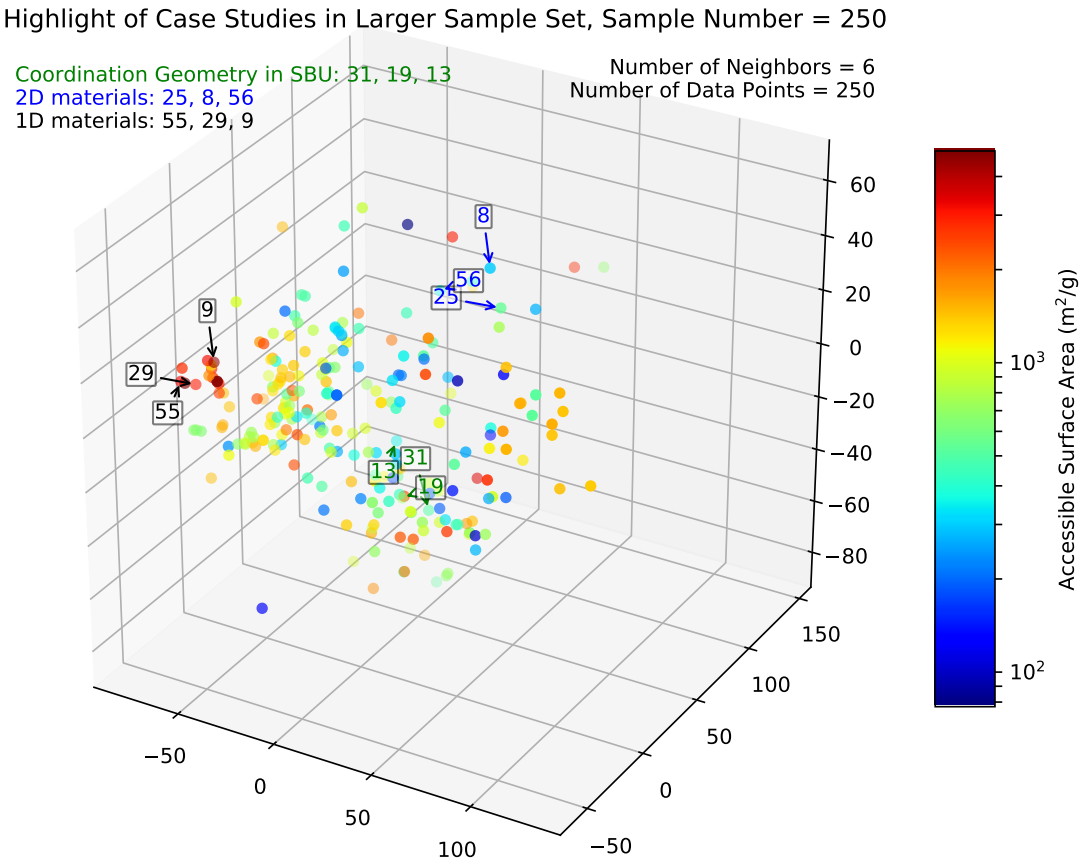

Figure S18: Network of 250 MOF structures show in 3D, without edges. The MOF structures in the case studies are annotated.

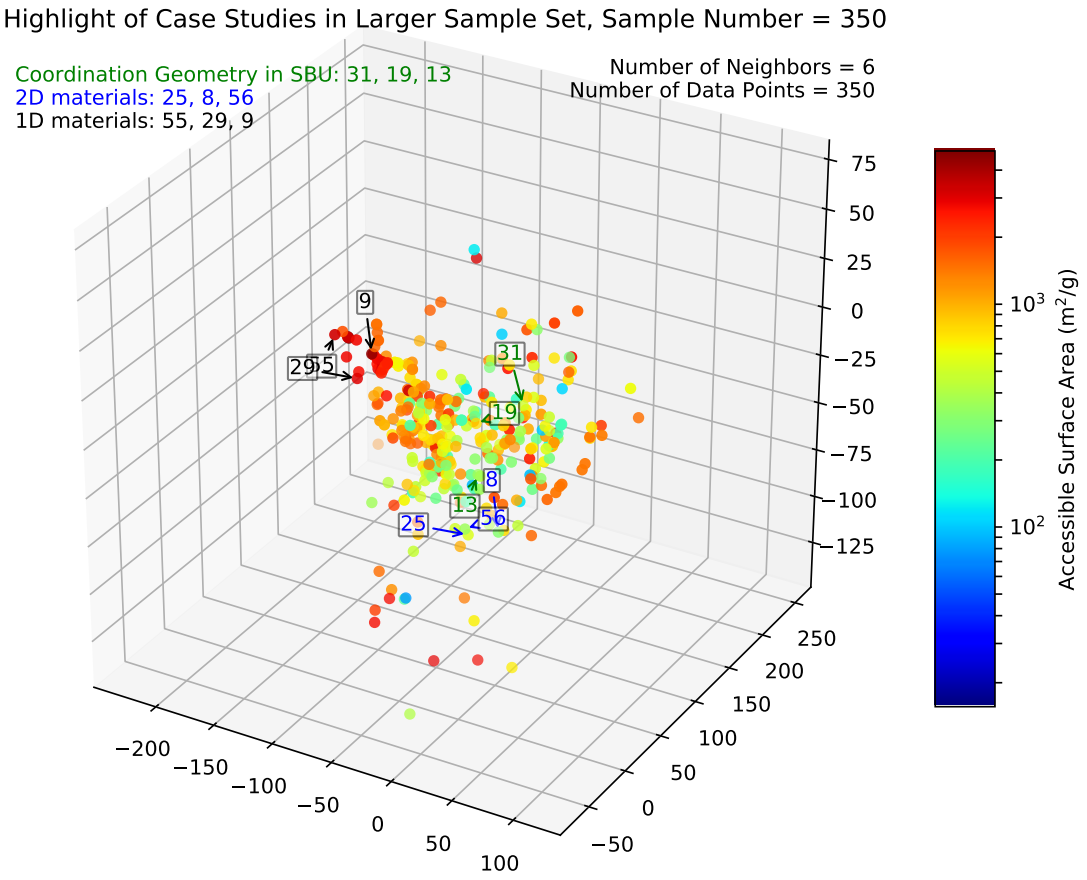

Figure S19: Network of 350 MOF structures show in 3D, without edges. The MOF structures in the case studies are annotated.

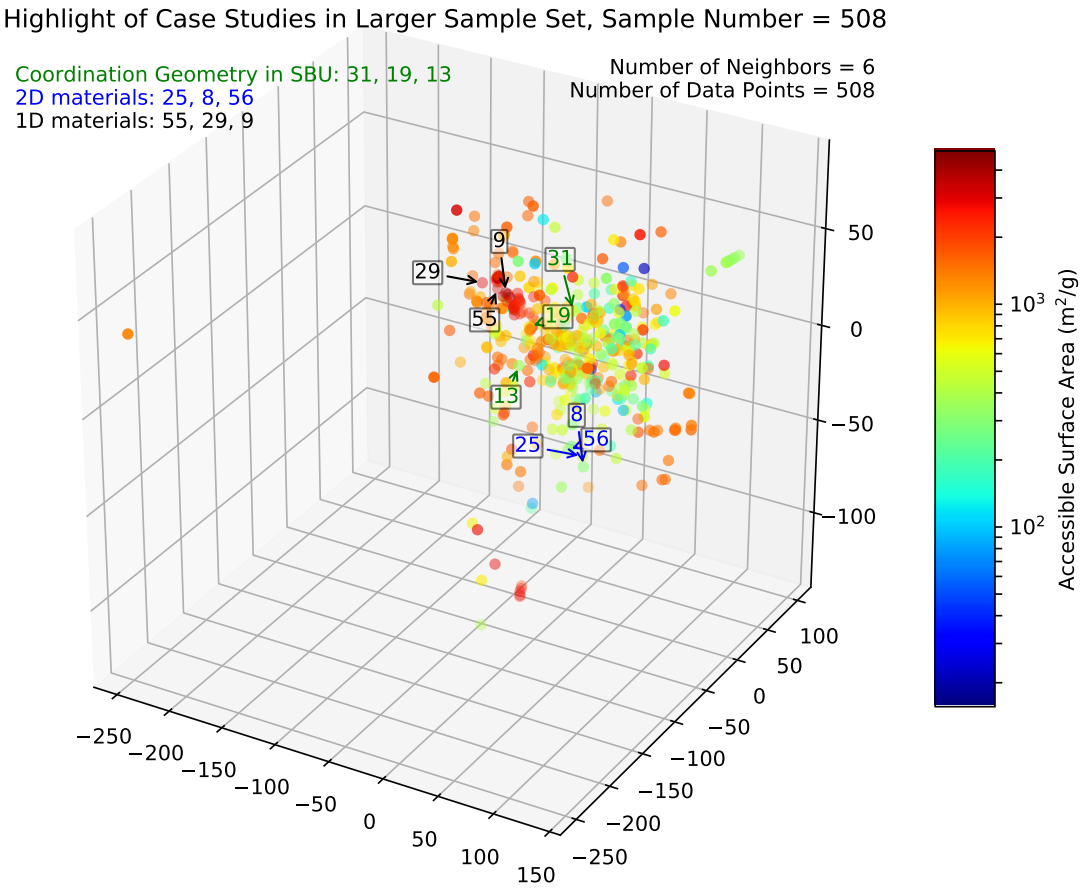

Figure S20: Network of 508 MOF structures show in 3D, without edges. The MOF structures in the case studies are annotated.

6.3 Effect of Increased Number of Samples in Dataset and Justification for Parameter Selection

The Isomap algorithm requires the number of nearest neighbors and the number of reconstruction dimensions as input parameters, and for different input datasets the optimal values of these parameters are likely to be different. To find the optimal parameter values for the different input datasets, we performed a reconstruction residual variance study. Specifically, we chose the values N = 50, 100, 150, 250 and 350 to be the number of input MOF structures for the input dataset. For each iteration, N MOF structures are selected from the 508 MOF structures randomly, the numbers of neighbors and dimensions of the reconstruction are varied, and reconstruction residual variances are then calculated based on the method discussed in Section 3 of this Supplementary Materials. This process is repeated 10 times and the average residual variance is calculated. Figure S21 shows the residual variance for different numbers of input structures. As can be seen, for the number of input

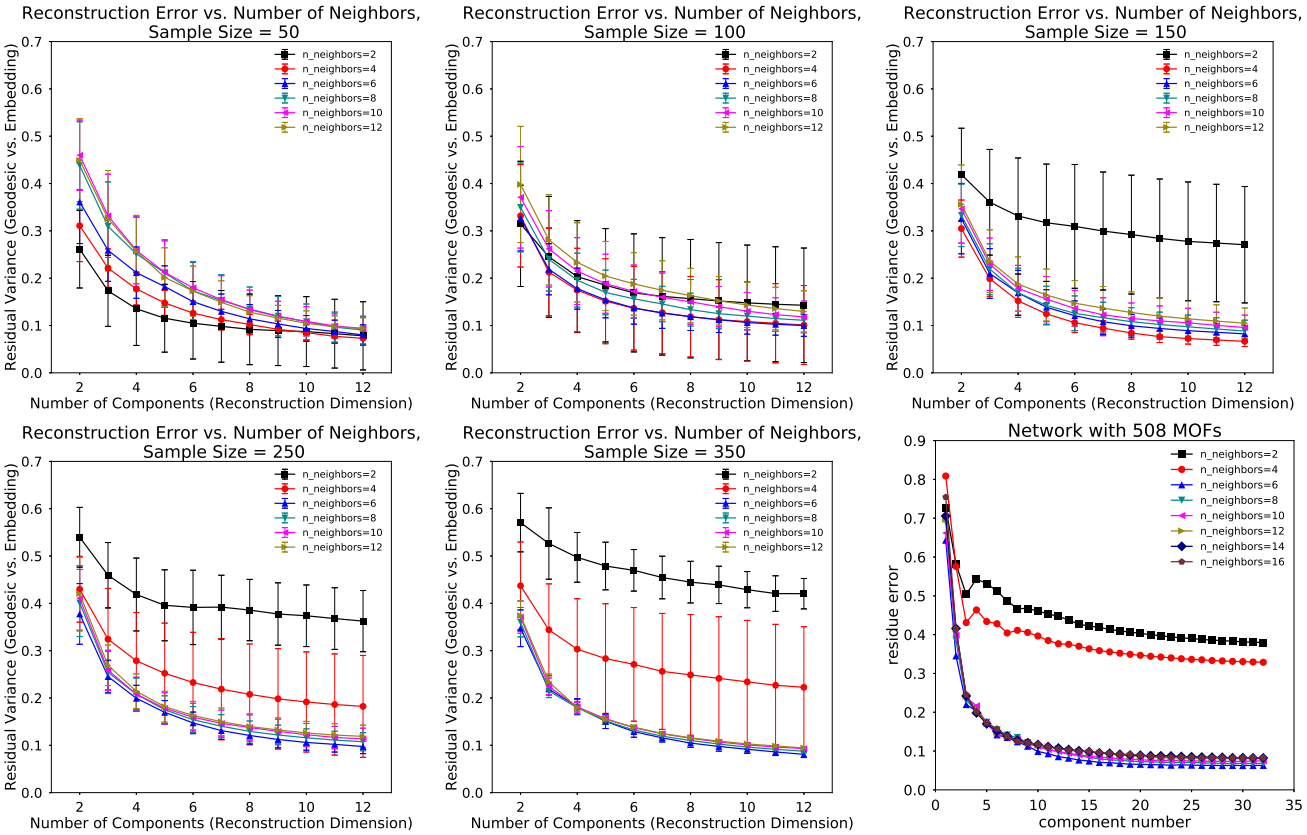

Figure S21: Reconstruction Residual Variances

MOF structures at 50, only 2 nearest neighbors are needed, while a larger number of nearest neighbors would result in an increase in the residual variance. Also in this case, 2 reconstruction dimensions has a residual variance at about .25. Starting from the input dataset with a size of 150 MOF structures, 4 neighbors are needed for the residual variance to be low. For a reconstruction dimension of 2, the residual variance is .3. The above observation can be explained by the fact that with more MOF structures added to the input dataset, the intrinsic dimension of the manifold represented by the graph network will also likely increase. This in turn will require a larger number of nearest neighbors for each data point and a larger number of reconstruction dimensions to preserve the manifold structure in the reconstruction. For the sample sizes of 150, 250, 350 and 508, the number of nearest neighbors are chosen to be 4, 6, 6, and 6.

The randomly generated indices for every iteration are available in the format of a spreadsheet in the online version.

7 Software

The Hirshfeld surfaces and the related properties are generated by Tonto (backend of CrystalExplorer[1]). The resolution for the surface calculation is set to "high", with 0.2 as the "desired_separation" of the plot grid. Python is used to read the data files generated by Tonto and Numpy is used to construct the 2D fingerprint plots. The option of normalization of the 2D histogram function in Numpy is set to be "True". Sci-kit Learn [6] package is used for the ISOMAP algorithm, and the neighborhood algorithm used is "kd_tree". For other input parameters, except for the number of components and number of nearest neighbors, the default settings are used. Additional packages used including: NetworkX[7], SciPy[8], Pandas[9] and Matplotlib[10].

References

[1] Mark A. Spackman and Dylan Jayatilaka. "Hirshfeld surface analysis". en. In: *CrystEngComm* 11.1 (Jan. 2009), pp. 19–32. ISSN: 1466-8033. DOI: 10.1039/B818330A. URL: http://pubs.rsc.org/en/content/articlelanding/2009/ce/b818330a (visited on 05/25/2018).

[2] C. R. Groom et al. "The Cambridge Structural Database". en. In: *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* 72.2 (Apr. 2016), pp. 171–179. ISSN: 2052-5206. DOI: 10.1107/S2052520616003954. URL: http://scripts.iucr.org/cgi-bin/paper?bm5086 (visited on 08/21/2018).

[3] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. "A Global Geometric Framework for Nonlinear Dimensionality Reduction". en. In: *Science* 290.5500 (Dec. 2000), pp. 2319–2323. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.290.5500.2319. URL: http://science.sciencemag.org/content/290/5500/2319 (visited on 02/08/2018).

[4] Mukund Balasubramanian and Eric L. Schwartz. "The Isomap Algorithm and Topological Stability". en. In: *Science* 295.5552 (Jan. 2002), pp. 7–7. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.295.5552.7a. URL: http://science.sciencemag.org/content/295/5552/7 (visited on 02/27/2018).

[5] William J. Cukierski and David J. Foran. "Using Betweenness Centrality to Identify Manifold Shortcuts". In: *Proceedings / IEEE International Conference on Data Mining. IEEE International Conference on Data Mining* 2008 (Dec. 2008), pp. 949–958. ISSN: 1550-4786. DOI: 10.1109/ICDMW.2008.39. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2895570/ (visited on 02/07/2018).

[6] Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (Oct. 2011), 28252830. URL: http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html (visited on 02/27/2018).

[7] Daniel A. Schult. "Exploring network structure, dynamics, and function using NetworkX". In: *In Proceedings of the 7th Python in Science Conference (SciPy.* 2008, pp. 11–15.

[8] Eric Jones, Travis Oliphant, and Pearu Peterson. "{SciPy}: Open source scientific tools for {Python}". In: (2001). URL: http://www.scipy.org (visited on 02/27/2018).

[9] W. McKinney. "Data structures for statistical computing in python". In: *Proceedings of the 9th Python in Science Conference* (2010), pp. 51–56.

[10] J. D. Hunter. "Matplotlib: A 2D Graphics Environment". In: *Computing in Science Engineering* 9.3 (May 2007), pp. 90–95. ISSN: 1521-9615. DOI: 10.1109/MCSE.2007.55.