The research of solvents sorption on Graphene Oxides

Grigory S. Ulyanov

December 31, 2025

1 Clustering methods

1.1 K-Means

K-Means is one of the most widely used partition-based clustering algorithms. It aims to divide a dataset into K clusters by minimizing the intra-cluster variance. The algorithm iteratively assigns each data point to the nearest cluster center and then updates the cluster centers as the mean of the assigned points. Due to its simplicity and computational efficiency, K-Means is especially suitable for large-scale datasets; however, it assumes roughly spherical and balanced clusters, which can limit its performance in more complex distributions.

1.2 K-Means++ Initialization Strategy

The performance of K-Means is sensitive to the initial selection of cluster centers, and poor initialization can lead to unfavorable local minima. To address this limitation, the K-Means++ initialization algorithm was proposed. Instead of selecting initial centers randomly, K-Means++ chooses each new center with probability proportional to the squared distance from the closest existing center, which increases the likelihood that centers are well separated before the iterative refinement begins. This approach significantly improves stability and typically leads to lower clustering error in practice [1].

1.3 Agglomerative Hierarchical Clustering

Agglomerative hierarchical clustering is a bottom-up approach that starts with each data point considered as an individual cluster. At each step, the two clusters with the highest similarity are merged until all points are grouped into a single cluster. The result of the process is often visualized using a dendrogram, which provides insight into the data's hierarchical structure and allows selecting the number of clusters by cutting the dendrogram at a chosen level. This method does not require specifying the number of clusters beforehand and is well-suited for identifying nested cluster structures [2].

1.4 Spectral Clustering

Spectral clustering is a graph-based clustering method that relies on the eigenstructure of matrices derived from pairwise similarities between data points. Instead of clustering directly in the original feature space, the data are first represented as a weighted graph, where nodes correspond to samples and edge weights encode similarity. A Laplacian matrix is then constructed, typically in the form $L = D - A$, or in its normalized variants such as $L_{\text{sym}} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ and $L_{\text{rw}} = D^{-1} A$, where A is the adjacency matrix and D is the diagonal degree matrix. By computing the eigenvectors associated with the smallest eigenvalues of the Laplacian, the data are embedded into a lower-dimensional space where the cluster structure becomes more distinct. Finally, a standard algorithm such as K-Means is applied in this spectral embedding

space. Due to its ability to capture complex, non-convex cluster shapes and exploit global structural properties in the data, spectral clustering is widely used in applications where traditional partition-based methods may fail [3].

1.5 DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm designed to identify clusters of arbitrary shape by locating regions of high point density. Unlike partition-based methods, DBSCAN does not require specifying the number of clusters in advance. Instead, it relies on two parameters: the neighborhood radius ε and the minimum number of points $MinPts$ required to form a dense region. A point is classified as a core point if at least $MinPts$ points fall within its ε-neighborhood. Clusters are formed by connecting core points that are density-reachable from one another, while points that do not belong to any cluster are labeled as noise. This property allows DBSCAN to effectively discover clusters of non-linear, non-convex shapes and to handle outliers naturally. However, its performance may be sensitive to the choice of ε and $MinPts$, particularly in datasets with varying density levels [4].

2 Substances Clustering

2.1 The baseline models

As the first step, the dataset of substances was loaded and the features such as dielectric constant and molar sorption were selected for further model development. The distribution of these features for each substance is presented in Figure 1. The K-Means and Hierarchical

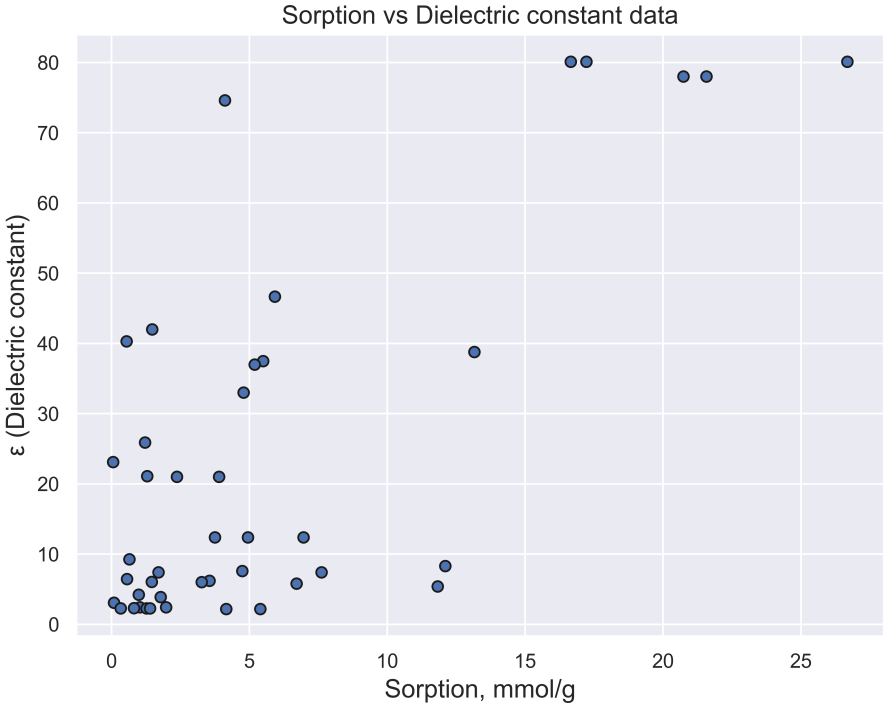
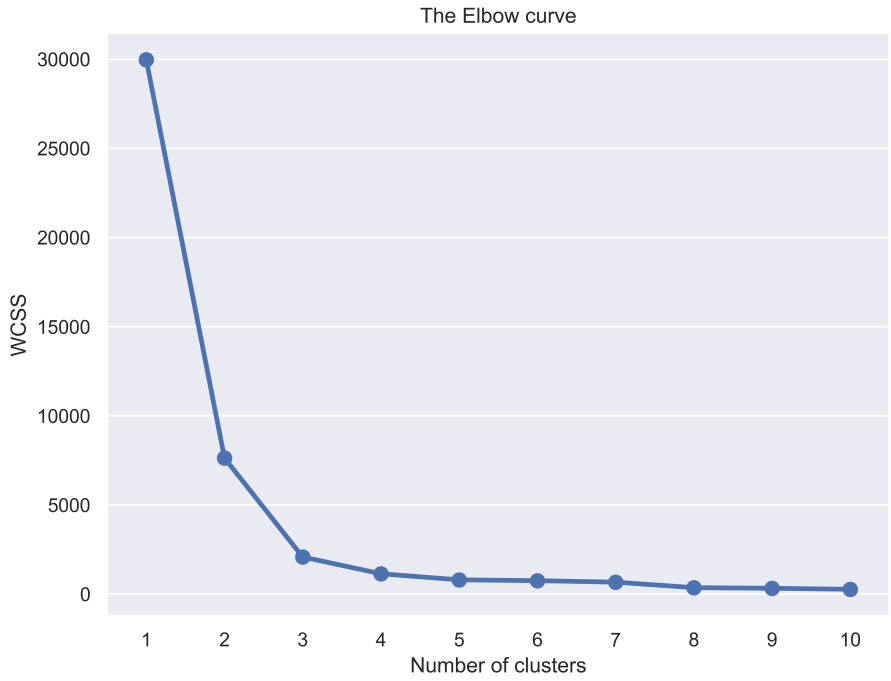


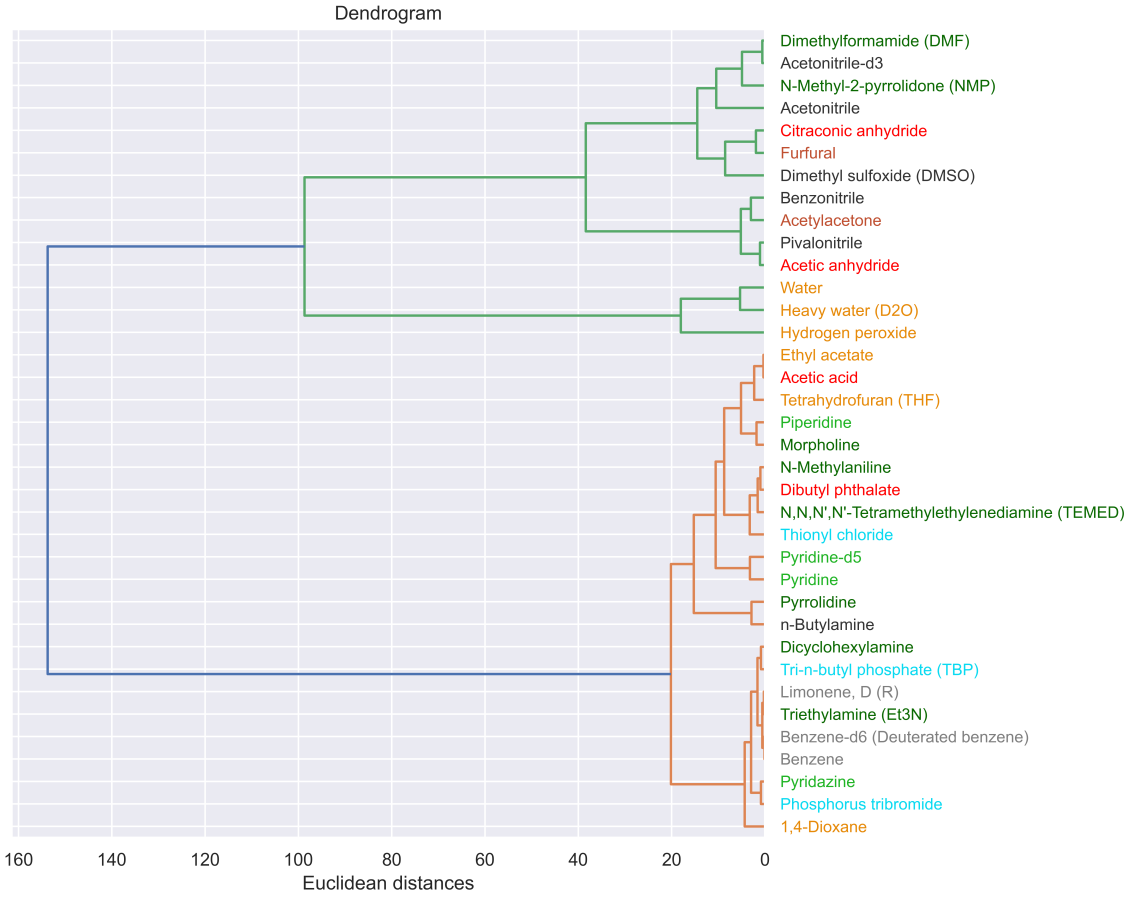
Figure 1: Distribution of dielectric constant and molar sorption across the substances.

clustering models are employed as baseline clustering approaches. The optimal number of clusters for K-Means is determined using the Elbow method, in which the within-cluster sum of squared errors is evaluated for different values of K. For the Hierarchical clustering model, the

number of clusters is inferred by visually inspecting the dendrogram and identifying the level at which the cluster merging distance increases sharply, indicating a natural separation in the data. The number of clusters was determined using Elbow curve and dendrogram (Figure 2). Using K-means, 4 clusters of compounds and their centroids were found (Figure 3).



(a) Elbow plot showing the within-cluster sum of squares for different values of K.



(b) Hierarchical clustering dendrogram illustrating cluster merging distances.

Figure 2: Visualization of cluster number selection for baseline clustering models.

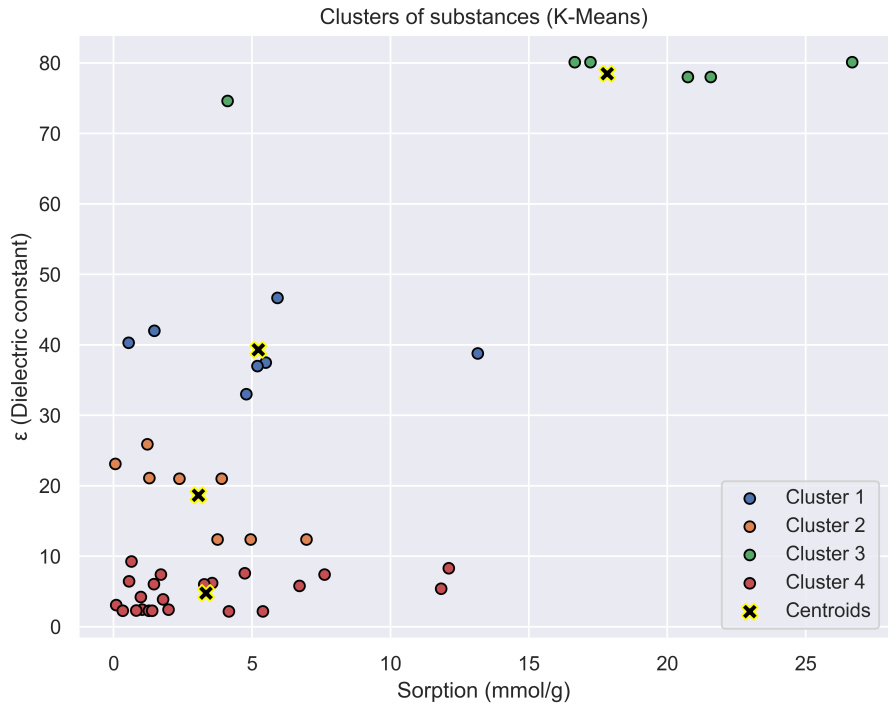


Figure 3: Clustering of compounds using K-Means with $K = 4$. Cluster centroids are shown in the feature space.

Using hierarchical agglomerative clustering, three distinct clusters of compounds were identified based on feature similarity (Figure 4).

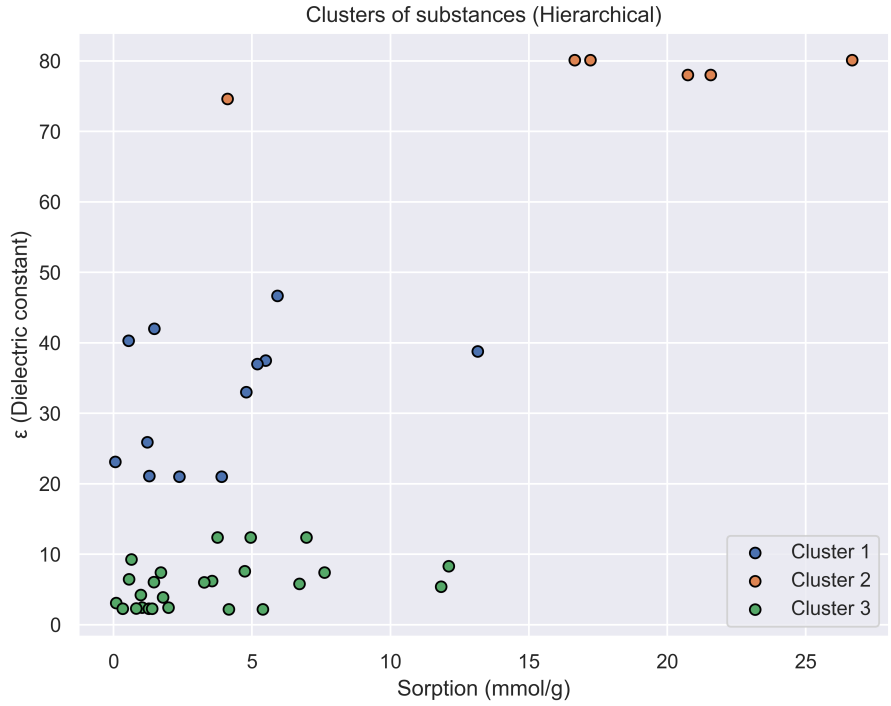


Figure 4: Hierarchical agglomerative clustering resulting in three clusters of compounds. The final cluster assignments are represented in the feature space.

To provide a clearer interpretation of the resulting cluster assignments, Tables 1 and 2

summarize the compounds grouped by the baseline K-Means and hierarchical agglomerative clustering models, respectively. For K-Means, four clusters were identified, while hierarchical clustering produced three clusters. The cluster compositions show consistent grouping trends, particularly in the separation of water- and peroxide-based solvents from organic amine and nitrile structures. Interactive visualizations of both clustering results are available in the project repository using Plotly-based dashboards, allowing the reader to explore cluster structures in a dynamic manner.

Table 1: Detailed cluster assignments obtained using K-Means clustering ($K = 4$): compounds, cluster labels and selected features.

Name	**Cluster**	ε (Dielectric constant)	**Sorption (mmol/g)**
N-Methyl-2-pyrrolidone (NMP)	1	33.00	4.791
Furfural	1	42.00	1.472
Citraconic anhydride	1	40.30	0.544
Dimethyl sulfoxide (DMSO)	1	46.68	5.919
Acetonitrile-d3	1	37.50	5.495
Acetonitrile	1	38.80	13.154
Dimethylformamide (DMF)	1	37.00	5.192
Acetonitrile	2	21.01	3.899
Pyridine	2	12.40	4.947
Pyridine	2	12.40	6.965
Pyridine-d5	2	12.40	3.752
Pivalonitrile	2	21.10	1.291
Benzonitrile	2	25.90	1.218
Acetylacetone	2	23.10	0.058
Acetic anhydride	2	21.00	2.373
Heavy water (D2O)	3	78.00	20.739
Heavy water (D2O)	3	78.00	21.568
Water	3	80.10	16.653
Water	3	80.10	26.679
Hydrogen peroxide	3	74.60	4.114
Water	3	80.10	17.220
Pyrrolidine	4	8.30	12.102
Pyridazine	4	4.22	0.983
Morpholine	4	7.40	7.618
Tetrahydrofuran (THF)	4	7.60	4.737
Piperidine	4	5.80	6.711
N-Methylaniline	4	6.06	1.455
1,4-Dioxane	4	2.20	5.392
Dicyclohexylamine	4	2.27	0.332
Dibutyl phthalate	4	6.44	0.561
Benzene-d6 (Deuterated benzene)	4	2.28	1.395
Benzene	4	2.28	1.271
N,N,N',N'-Tetramethylethylenediamine (TEMED)	4	7.40	1.705
Phosphorus tribromide	4	3.90	1.782
n-Butylamine	4	5.40	11.826
Triethylamine (Et_3N)	4	2.42	1.027
Triethylamine (Et_3N)	4	2.42	1.980
1,4-Dioxane	4	2.20	4.161
Tri-n-butyl phosphate (TBP)	4	3.09	0.091
Thionyl chloride	4	9.25	0.644
Acetic acid	4	6.20	3.553
Limonene, D (R)	4	2.30	0.814
Ethyl acetate	4	6.02	3.270

Table 2: Cluster assignments obtained using Hierarchical Clustering.

Name	**Cluster**	ε	**Sorption (mmol/g)**
Acetonitrile-d3	1	37.50	5.495
Acetic anhydride	1	21.00	2.373
N-Methyl-2-pyrrolidone (NMP)	1	33.00	4.791
Acetonitrile	1	38.80	13.154
Acetylacetone	1	23.10	0.058
Furfural	1	42.00	1.472
Citraconic anhydride	1	40.30	0.544
Pivalonitrile	1	21.10	1.291
Benzonitrile	1	25.90	1.218
Acetonitrile	1	21.01	3.899
Dimethylformamide (DMF)	1	37.00	5.192
Dimethyl sulfoxide (DMSO)	1	46.68	5.919
Water	2	80.10	26.679
Water	2	80.10	17.220
Heavy water (D2O)	2	78.00	20.739
Heavy water (D2O)	2	78.00	21.568
Water	2	80.10	16.653
Hydrogen peroxide	2	74.60	4.114
Piperidine	3	5.80	6.711
Pyrrolidine	3	8.30	12.102
Pyridine	3	12.40	4.947
Pyridine	3	12.40	6.965
Pyridine-d5	3	12.40	3.752
Morpholine	3	7.40	7.618
Tetrahydrofuran (THF)	3	7.60	4.737
Pyridazine	3	4.22	0.983
N-Methylaniline	3	6.06	1.455
1,4-Dioxane	3	2.20	5.392
Dicyclohexylamine	3	2.27	0.332
Dibutyl phthalate	3	6.44	0.561
Benzene-d6	3	2.28	1.395
Benzene	3	2.28	1.271
N,N,N',N'-Tetramethylethylenediamine (TEMED)	3	7.40	1.705
Phosphorus tribromide	3	3.90	1.782
n-Butylamine	3	5.40	11.826
Triethylamine (Et_3N)	3	2.42	1.027
Triethylamine (Et_3N)	3	2.42	1.980
1,4-Dioxane	3	2.20	4.161
Tri-n-butyl phosphate (TBP)	3	3.09	0.091
Thionyl chloride	3	9.25	0.644
Acetic acid	3	6.20	3.553
Limonene (D, R)	3	2.30	0.814
Ethyl acetate	3	6.02	3.270

2.2 Other models

Spectral clustering and DBSCAN were employed as non-baseline clustering approaches. For spectral clustering, an adjacency matrix A was constructed using a radial basis kernel of the

form

$$A_{ij} = \exp\Big(- \frac{\|x_i - x_j\|^2}{2\sigma^2} \Big),$$

which encodes pairwise similarity between samples. Here, x_i and x_j denote the feature vectors of the i-th and j-th data points in the original dataset, and $\|x_i - x_j\|$ represents the Euclidean distance between them in the feature space. The parameter σ controls the width of the Gaussian kernel and thus the notion of locality: small values of σ make the similarity decay sharply with distance, while larger values treat more distant points as similar.The scale parameter controls how rapidly the affinity decays with Euclidean distance. A common practical choice is $\sigma = 1$, which simplifies the kernel and assumes that the raw feature space already reflects an appropriate notion of distance. From A the degree matrix D and a graph Laplacian (e.g. $L = D - A$ or the normalized variant $L_{\text{sym}} = D^{-1/2} A D^{-1/2}$) were formed and the eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ of the Laplacian were computed. To select the number of clusters we applied the eigengap heuristic: the consecutive differences $\Delta_k = \lambda_{k+1} - \lambda_k$ were evaluated and plotted (Figure 5). The heuristic prescribes choosing K at the index where Δ_k attains a pronounced maximum, since a large gap indicates a natural separation between the k-dimensional low-energy subspace and the remaining spectrum. After embedding the data in the space spanned by the first K eigenvectors, a K-Means algorithm is applied to obtain the final clustering.

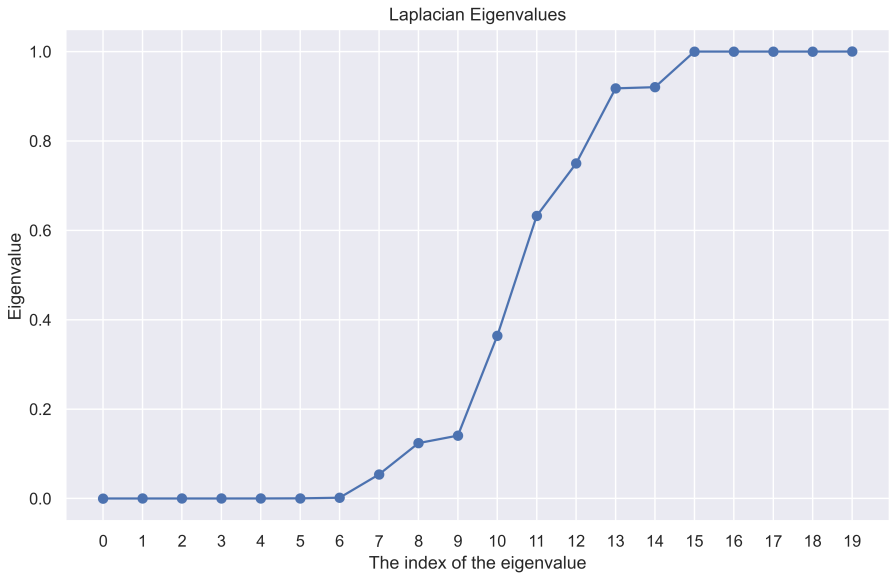


Figure 5: Eigenvalue index plot for the Laplacian constructed from the affinity matrix A. The plot shows eigenvalue magnitudes and highlights the location of the largest consecutive gap.

The eigengap analysis produced the following numeric results (differences between consecutive eigenvalues and the corresponding candidate numbers of clusters). The largest gap is shown in bold and suggests the optimal choice of K.

Table 3: Eigengap analysis: consecutive eigenvalue differences and suggested numbers of clusters.

Number of clusters	Difference between eigenvalues
2	6.875757×10^{-16}
3	1.925212×10^{-14}
4	7.640218×10^{-7}
5	6.200000×10^{-5}
6	2.730000×10^{-4}
7	1.400000×10^{-3}
8	5.213400×10^{-2}
9	$\mathbf{7.002400 \times 10^{-2}}$
10	1.682100×10^{-2}

In this analysis the maximum eigengap occurs at the entry corresponding to $\Delta_9 = 0.070024$, which indicates $\mathbf{K = 9}$ as the most natural partitioning according to the eigengap heuristic.

DBSCAN was also applied as a complementary method. DBSCAN does not require a priori specification of K; instead it uses a neighbourhood radius ε and a minimum sample count $MinPts$ to identify dense regions. Points with at least $MinPts$ neighbours within ε are considered core points and clusters are formed by density-reachability from these cores, while sparse points are labeled as noise.

2.2.1 Spectral clustering results

The spectral clustering procedure described above was executed using the affinity matrix based on Gaussian kernel. The clusterization plot is presented in Figure 6.

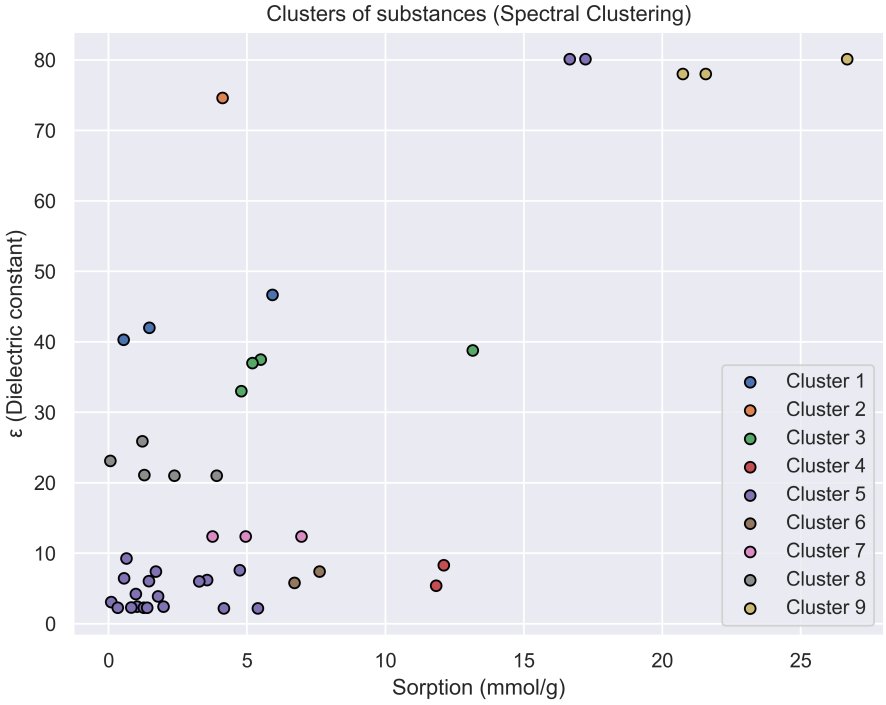


Figure 6: Spectral clustering resulting in nine clusters of compounds. The final cluster assignments are presented in the feature space.

The clustering results obtained using spectral clustering are summarized in Table 6. Unlike the other methods considered in this study, spectral clustering produced a substantially larger

number of clusters. This behaviour is explained by the fact that spectral clustering operates directly on the eigenstructure of the graph Laplacian constructed from the affinity matrix. The algorithm is sensitive to fine variations in local similarity, and the eigengap heuristic tends to identify multiple natural separations whenever the spectrum exhibits several comparable gaps. As a result, spectral clustering may partition the dataset into a larger number of coherent but smaller clusters, particularly when local neighbourhood structures differ significantly.

In contrast, methods such as K-Means or hierarchical clustering impose stronger global assumptions on cluster geometry, often leading to fewer, more aggregated clusters. Density-based approaches like DBSCAN, on the other hand, group points according to density connectivity and typically yield compact clusters while treating low-density regions as noise.

Table 4: Cluster assignments obtained using Spectral Clustering.

Name	**Cluster**	ε	**Sorption (mmol/g)**
Dimethyl sulfoxide (DMSO)	1	46.68	5.919
Citraconic anhydride	1	40.30	0.544
Furfural	1	42.00	1.472
Hydrogen peroxide	2	74.60	4.114
Acetonitrile-d3	3	37.50	5.495
N-Methyl-2-pyrrolidone (NMP)	3	33.00	4.791
Dimethylformamide (DMF)	3	37.00	5.192
Acetonitrile	3	38.80	13.154
Pyrrolidine	4	8.30	12.102
n-Butylamine	4	5.40	11.826
N-Methylaniline	5	6.06	1.455
Benzene-d6 (Deuterated benzene)	5	2.28	1.395
Limonene (D,R)	5	2.30	0.814
Tetrahydrofuran (THF)	5	7.60	4.737
Dicyclohexylamine	5	2.27	0.332
Dibutyl phthalate	5	6.44	0.561
Water	5	80.10	16.653
Pyridazine	5	4.22	0.983
1,4-Dioxane	5	2.20	5.392
Water	5	80.10	17.220
Ethyl acetate	5	6.02	3.270
N,N,N',N'-Tetramethylethylenediamine (TEMED)	5	7.40	1.705
Acetic acid	5	6.20	3.553
Phosphorus tribromide	5	3.90	1.782
Triethylamine (Et_3N)	5	2.42	1.027
Triethylamine (Et_3N)	5	2.42	1.980
Thionyl chloride	5	9.25	0.644
1,4-Dioxane	5	2.20	4.161
Benzene	5	2.28	1.271
Tri-n-butyl phosphate (TBP)	5	3.09	0.091
Piperidine	6	5.80	6.711
Morpholine	6	7.40	7.618
Pyridine	7	12.40	4.947
Pyridine	7	12.40	6.965
Pyridine-d5	7	12.40	3.752
Acetylacetone	8	23.10	0.058
Acetonitrile	8	21.01	3.899
Acetic anhydride	8	21.00	2.373
Pivalonitrile	8	21.10	1.291
Benzonitrile	8	25.90	1.218
Water	9	80.10	26.679
Heavy water (D_2O)	9	78.00	20.739
Heavy water (D_2O)	9	78.00	21.568

2.2.2 DBSCAN clustering results

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was applied as a complementary, non-parametric clustering method. In this work, the neighbourhood radius was

set to $\varepsilon = 3$ and the minimum number of points to MinPts $= 6$. These values were chosen empirically: among a wide range of tested parameter combinations, $\varepsilon = 3$ and MinPts $= 6$ produced the smallest noise component (i.e., the smallest cluster assigned to outliers) while simultaneously yielding a clustering structure that remained interpretable with respect to the underlying chemistry of the compounds.

The resulting DBSCAN partitioning is presented in Figure 7 and Table 5, which illustrates the distribution of the four identified clusters in the original feature space. One of these clusters corresponds to noise points, shown separately in the figure as the "noise cluster". This component captures samples that do not belong to any sufficiently dense region and therefore cannot be reliably assigned to a chemically meaningful group.

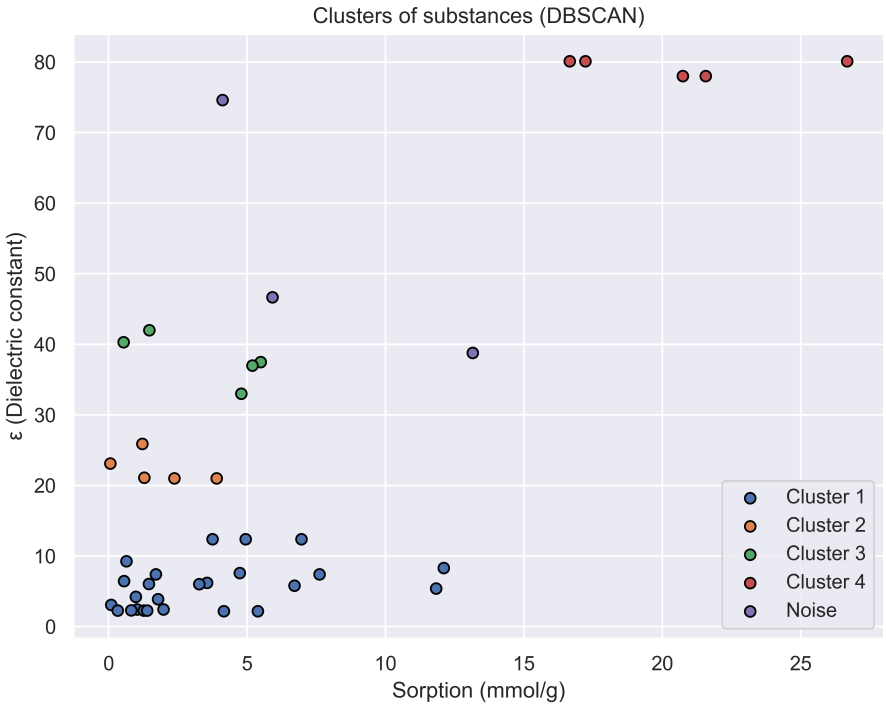


Figure 7: DBSCAN clustering resulting in four clusters of compounds. The final cluster assignments are presented in the feature space. One of the clusters corresponds to the noise points identified by DBSCAN.

A key advantage of DBSCAN in this application is its ability to identify arbitrarily shaped clusters and to explicitly detect outliers without requiring the number of clusters to be specified in advance. This is in contrast to spectral clustering, which relies on selecting the number of clusters using spectral properties of the graph Laplacian. As a result, spectral clustering may partition the data into a relatively large number of clusters when the affinity structure contains several moderately separated subgroups, whereas DBSCAN can merge such regions when they form a continuous high-density manifold. Hence, DBSCAN provides a complementary perspective by emphasizing density-based structure rather than global similarity patterns.

Table 5: Cluster assignments obtained using DBSCAN. Cluster 0 corresponds to noise points.

Name	**Cluster**	**ε (Dielectric constant)**	**Sorption (mmol/g)**
Hydrogen peroxide	0	74.60	4.114
Dimethyl sulfoxide (DMSO)	0	46.68	5.919
Acetonitrile	0	38.80	13.154
Benzene-d6 (Deuterated benzene)	1	2.28	1.395
Dibutyl phthalate	1	6.44	0.561
Dicyclohexylamine	1	2.27	0.332
Limonene, D (R)	1	2.30	0.814
N-Methylaniline	1	6.06	1.455
Morpholine	1	7.40	7.618
1,4-Dioxane	1	2.20	5.392
Pyridazine	1	4.22	0.983
Pyridine	1	12.40	4.947
Pyridine	1	12.40	6.965
Pyridine-d5	1	12.40	3.752
Pyrrolidine	1	8.30	12.102
Tetrahydrofuran (THF)	1	7.60	4.737
Piperidine	1	5.80	6.711
Benzene	1	2.28	1.271
Ethyl acetate	1	6.02	3.270
Phosphorus tribromide	1	3.90	1.782
Acetic acid	1	6.20	3.553
Thionyl chloride	1	9.25	0.644
n-Butylamine	1	5.40	11.826
Triethylamine (Et_3N)	1	2.42	1.027
Triethylamine (Et_3N)	1	2.42	1.980
Tri-n-butyl phosphate (TBP)	1	3.09	0.091
N,N,N',N'-Tetramethylethylenediamine (TEMED)	1	7.40	1.705
1,4-Dioxane	1	2.20	4.161
Acetic anhydride	2	21.00	2.373
Acetylacetone	2	23.10	0.058
Pivalonitrile	2	21.10	1.291
Acetonitrile	2	21.01	3.899
Benzonitrile	2	25.90	1.218
Citraconic anhydride	3	40.30	0.544
Furfural	3	42.00	1.472
Dimethylformamide (DMF)	3	37.00	5.192
N-Methyl-2-pyrrolidone (NMP)	3	33.00	4.791
Acetonitrile-d3	3	37.50	5.495
Water	4	80.10	26.679
Heavy water (D_2O)	4	78.00	20.739
Heavy water (D_2O)	4	78.00	21.568
Water	4	80.10	16.653
Water	4	80.10	17.220

2.3 Quantitative comparison of clustering methods

The so-called internal methods were used to assess the quality of clustering. These measures assess the quality of the cluster structure based only directly on it, without using external information. Three widely used metrics were used in our work: Silhouette Score, Calinski–Harabasz and Davies–Bouldin indexes. The silhouette value shows how similar an object is to its cluster compared to other clusters. This parameter ranges from -1 to 1, so the closer this score is to 1, the better. In the Calinski–Harabasz index, compactness is based on the distance from cluster points to their centroids, and separability is based on the distance from cluster centroids to the global centroid. An increase in this index gives a higher clustering quality. The Davies–Bouldin index calculates compactness as the distance from cluster objects to their centroids, and separability as the distance between the centroids. This index should be minimized for clusterization growth.

Below are the brief mathematical definitions of these metrics.

- **Silhouette score (per sample)**:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i),\, b(i)\}},$$

where $a(i)$ is the average distance from sample i to all other points in the same cluster and $b(i)$ is the minimum average distance from i to points in any other single cluster. The overall silhouette score is the mean of $s(i)$ across all samples.

- **Calinski–Harabasz index**:

$$\mathrm{CH} = \frac{\mathrm{tr}(B_k)/(k-1)}{\mathrm{tr}(W_k)/(n-k)},$$

where n is the number of samples, k the number of clusters, $\mathrm{tr}(B_k)$ is the between-cluster dispersion and $\mathrm{tr}(W_k)$ is the within-cluster dispersion. Larger CH indicates more compact and better separated clusters.

- **Davies–Bouldin index**:

$$\mathrm{DB} = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i} \frac{S_i + S_j}{M_{ij}},$$

where S_i is a measure of dispersion and M_{ij} is the distance between centroids of clusters i and j. Lower DB is better (smaller intra-cluster scatter and larger inter-cluster separation).

Table 6: Internal clustering metrics computed for each method. Higher Silhouette and Calinski–Harabasz are better; lower Davies–Bouldin is better.

Method	**Silhouette**	**Calinski–Harabasz**	**Davies–Bouldin**
K-Means	0.61	328.85	0.50
Hierarchical	**0.70**	**285.95**	**0.43**
Spectral	0.06	7.30	2.70
DBSCAN	0.60	185.67	0.76

Thus, according to the obtained metrics, the highest quality clustering method is hierarchical clustering. K-Means is also a qualitative method, but with slightly worse metrics. The metrics of the DBSCAN method can be improved by tuning hyperparameters. Spectral clustering showed unsatisfactory results, since during the calculation of the adjacency matrix, most of the values in the matrix were zero, which indicates a high value of dissimilarity and the output of an abnormal number of clusters when calculating the eigenvalues of the Laplacian (Figure 8).

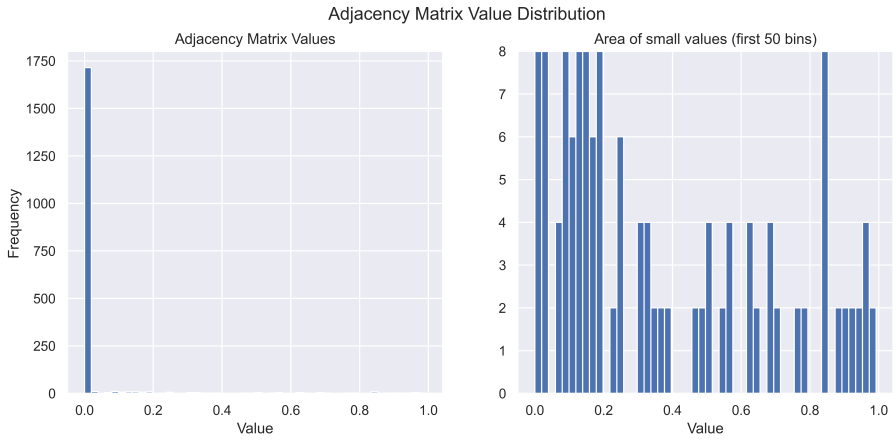


Figure 8: Distribution of adjacency matrix values at the parameter $\sigma = 1$.

In the future, the sigma value will be tuned in the adjacency matrix, since this clustering method is a powerful tool for finding relationships in complex data, which is ours. Despite the

metrics, the next step of the work is to find linear correlations between sorption and permittivity for compounds within clusters from all the methods previously used.

3 RANSAC Regression

RAndom SAmple Consensus regression or RANSAC regression is a robust regression method based on building a model, such as linear regression, on a set of random subsamples of data and evaluating their quality to build a so-called consensus set of points consistent with the resulting model. These points are called inliers and the best choice of inliers is the regression output. To set the robustness of the model, a threshold value is used, which is calculated from the distribution of data or selected by the user based on his considerations [5]. In our work, the threshold parameter for RANSAC is 1, since this value has a physical meaning, namely, a point is considered to be inlier if it deviates from the line by less than 1 unit dielectric constant. Choosing this parameter gives us a hard regression threshold, which is extremely important in the task. The paper also made attempts to calculate this threshold value based on the median absolute deviation, which were not crowned with satisfactory and interpretable results.

Below is a visualization of this algorithm for all previously used clustering methods: for K-Means (Fig. 9), for Hierarchical clustering (Fig. 10), for Spectral clustering (Fig. 11) and for DBSCAN (Fig. 12). Note that linear regression requires at least 3 points in order to be constructed, so the algorithm was not run through small clusters containing fewer points.

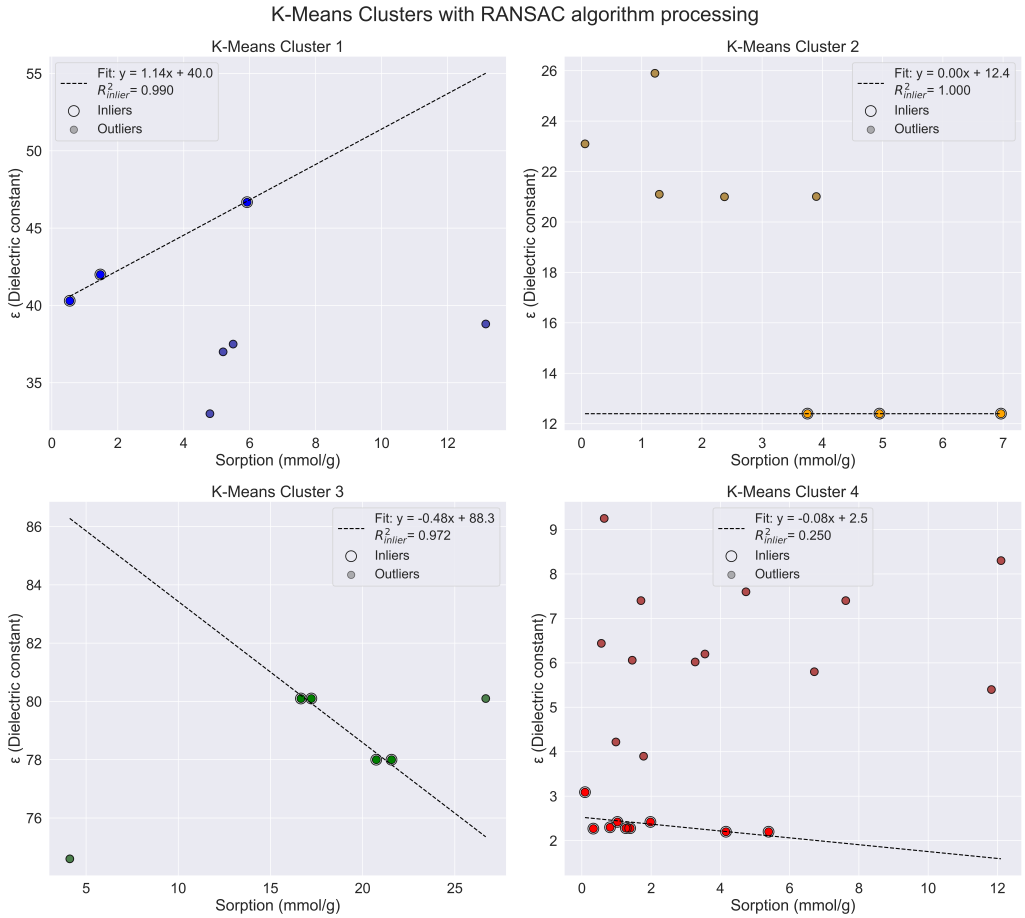


Figure 9: RANSAC Regression within the K-Means clusters.

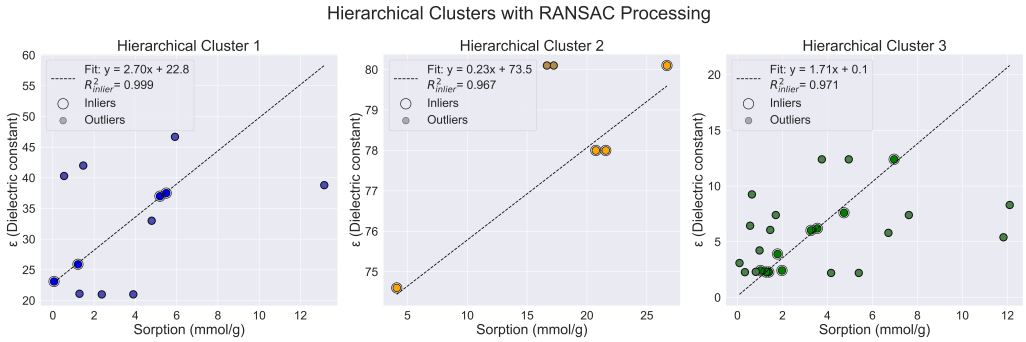


Figure 10: RANSAC Regression within the Hierarchical clusters.

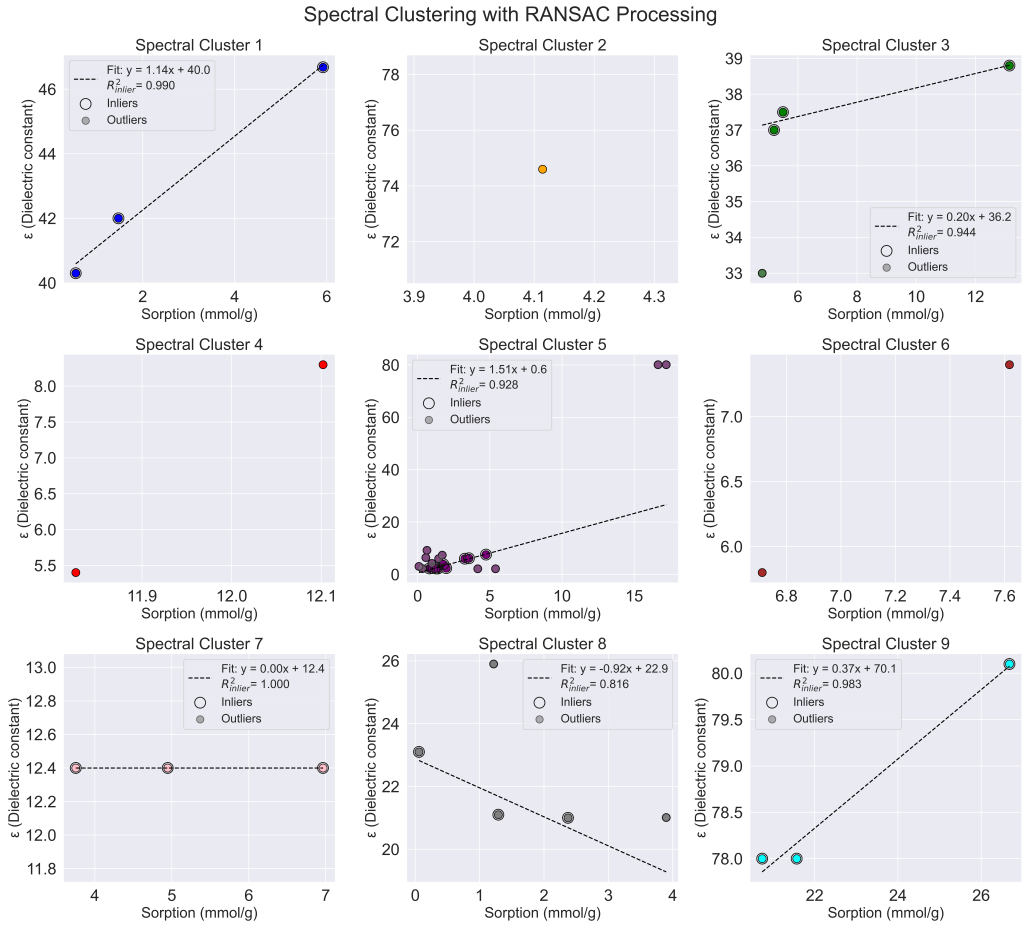


Figure 11: RANSAC Regression within the Spectral clusters.

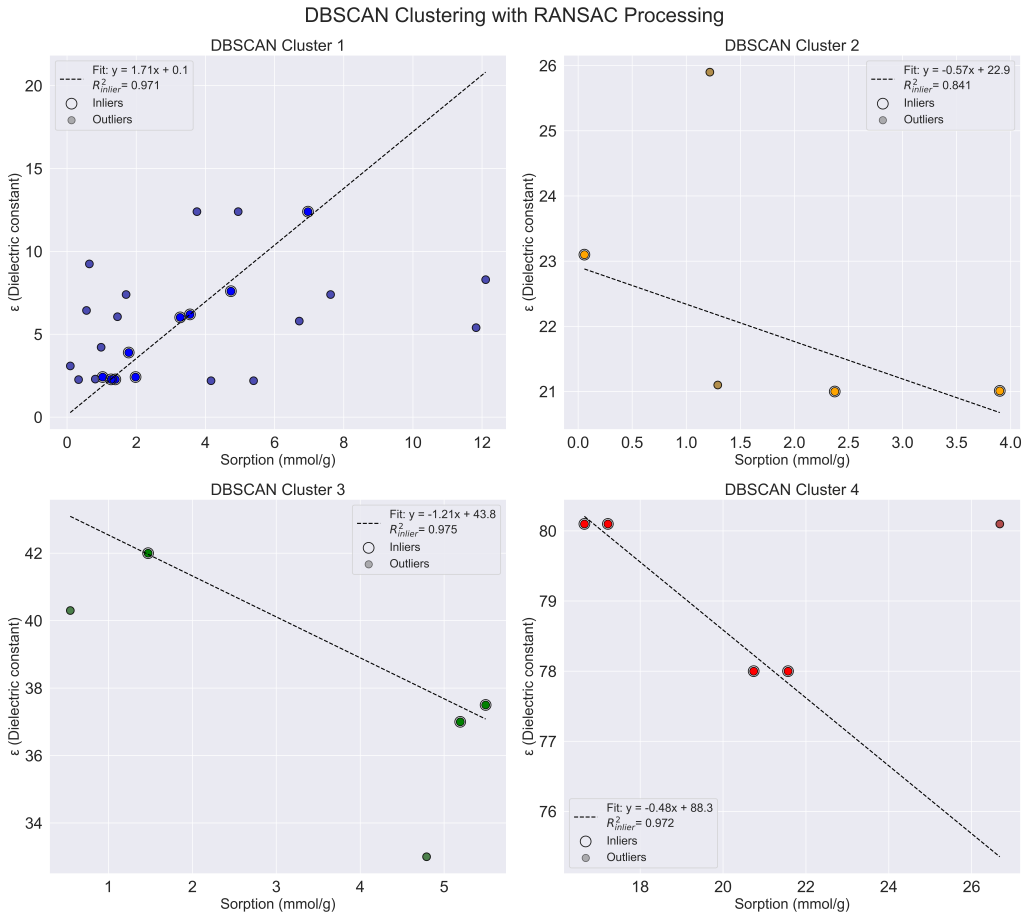


Figure 12: RANSAC Regression within the DBSCAN clusters.

Regression equations for the initial data with determination coefficients were obtained for each cluster processed using RANSAC. One can see that some clusters contain more linear correlations than one. We also observe the presence of negative correlations in some cases and low coefficients of determination. For a more thorough analysis of linear correlations and obtaining interpretable results, it was decided to tighten and supplement RANSAC to find strictly positive correlations that can be interpreted, as opposed to negative ones. Solutions have also been found for finding several linear correlations for large clusters. This solution is described in the following paragraph.

4 Sequantial RANSAC

The sequential RANSAC method is necessary for constructing several linear correlations without neglecting other robust methods. The essence of this method is simple and consists in masking the first group of inliers from the second, the second from the third, and so on. The method was applied to three clustering methods, excluding the spectral one. 1 was also chosen as the threshold value, a limit on the positive slope was introduced, and in addition, a boundary coefficient of determination equal to 0.65 was introduced. Let's consider the results of sequential RANSAC for the above clustering methods in more detail.

4.1 Sequantial RANSAC for K-Means

Clusters 1 and 4 in K-Means need the sequential RANSAC method, when the rest of the clusters are small for such a method and do not contain positive correlations, based on the previous

paragraph. The graph for clusters 1 and 4 in K-Means is shown in the Figure 13. The Table 7 shows the inliers, numbered in this plot.

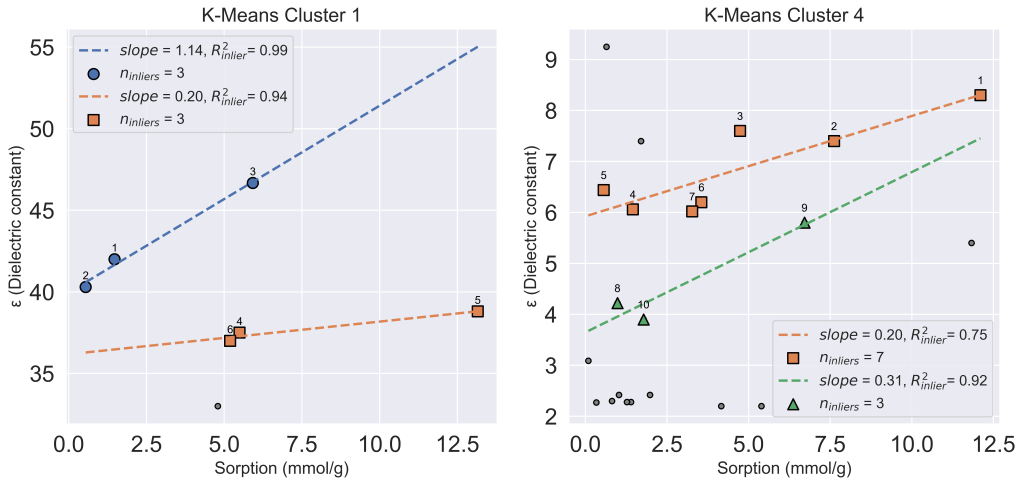


Figure 13: Sequential RANSAC Regression within the K-Means clusters 1 and 4.

Table 7: Inliers found by sequential RANSAC in K-Means clusters 1 and 4. Each row lists the cluster, RANSAC iteration, inlier index and the corresponding compound with its feature values.

Cluster	Inlier number	Local index	Name
1	1	1	Furfural
1	2	2	Citraconic anhydride
1	3	3	Dimethyl sulfoxide (DMSO)
1	4	4	Acetonitrile-d3
1	5	5	Acetonitrile
1	6	6	Dimethylformamide (DMF)
4	1	21	Pyrrolidine
4	2	23	Morpholine
4	3	24	Tetrahydrofuran (THF)
4	4	26	N-Methylaniline
4	5	29	Dibutyl phthalate
4	6	40	Acetic acid
4	7	42	Ethyl acetate
4	8	22	Pyridazine
4	9	25	Piperidine
4	10	33	Phosphorus tribromide

The numbered annotations in Figure 13 correspond to the rows of Table 7 (row order 1–16). In the figure the marked compounds are, in order: (1) Furfural; (2) Citraconic anhydride; (3) Dimethyl sulfoxide (DMSO); (4) Acetonitrile-d3; (5) Acetonitrile; (6) Dimethylformamide (DMF); (7) Pyrrolidine; (8) Morpholine; (9) Tetrahydrofuran (THF); (10) N-Methylaniline; (11) Dibutyl phthalate; (12) Acetic acid; (13) Ethyl acetate; (14) Pyridazine; (15) Piperidine; (16) Phosphorus tribromide. These labels match the inlier numbering shown on the plot and allow the reader to cross-reference each annotated point with its full entry in Table 7.

4.2 Sequantial RANSAC for Hierarchical clustering

Separate graphs were selected for each of the Hierarchical clusters for qualitative display. The graph for the Hierarchical cluster series is shown in Figures 14, 15, 16. Table 8 lists the inliers

detected by sequential RANSAC in the Hierarchical clustering result; the numbering in the table corresponds to the numbered annotations in the figure.

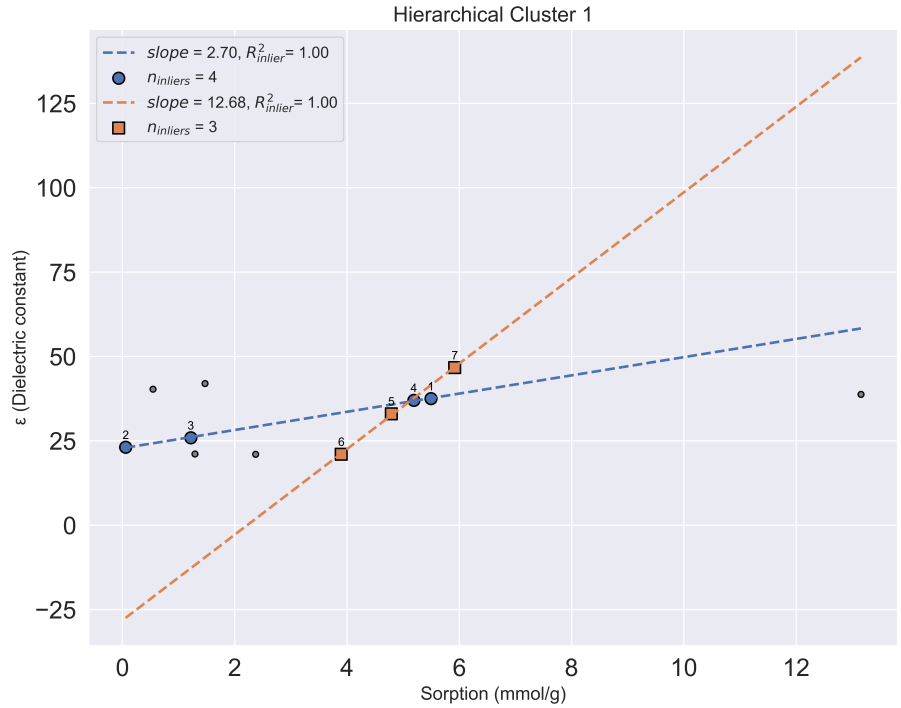


Figure 14: Sequential RANSAC Regression within the Hierarchical cluster 1.

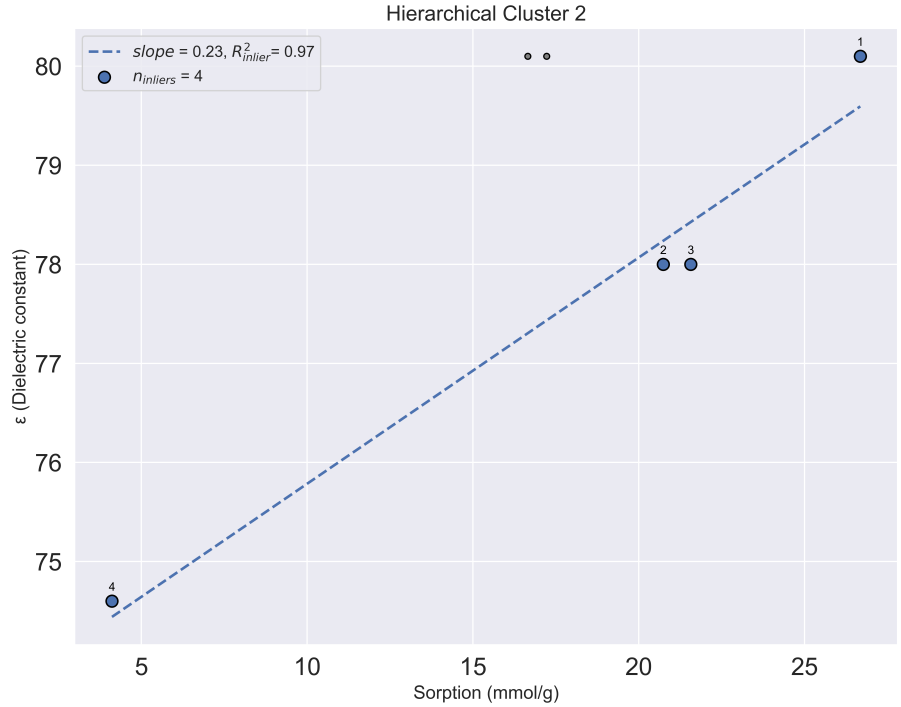


Figure 15: Sequential RANSAC Regression within the Hierarchical cluster 2.

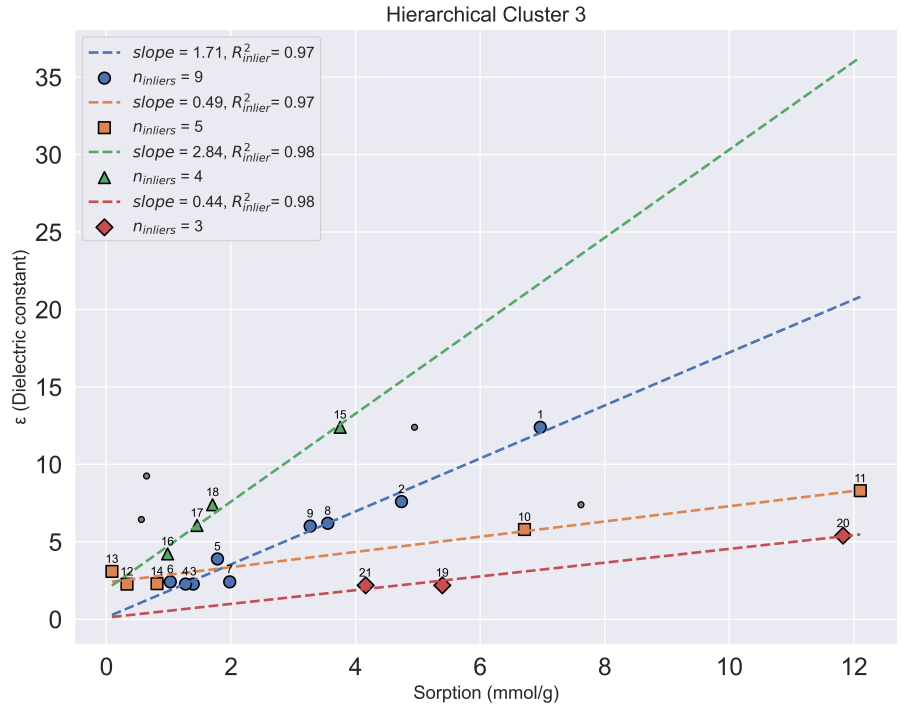


Figure 16: Sequential RANSAC Regression within the Hierarchical cluster 3.

The marked compounds in the figure are, in order: (1) Acetonitrile-d3; (2) Acetylacetone; (3) Benzonitrile; (4) Dimethylformamide (DMF); (5) N-Methyl-2-pyrrolidone (NMP); (6) Acetonitrile; (7) Dimethyl sulfoxide (DMSO); (8) Water; (9) Heavy water (D2O); (10) Heavy water (D2O); (11) Hydrogen peroxide; (12) Pyridine; (13) Tetrahydrofuran (THF); (14) Benzene-d6 (Deuterated benzene); (15) Benzene; (16) Phosphorus tribromide; (17) Triethylamine (Et_3N); (18) Triethylamine (Et_3N); (19) Acetic acid; (20) Ethyl acetate; (21) Piperidine; (22) Pyrrolidine; (23) Dicyclohexylamine; (24) Tri-n-butyl phosphate (TBP); (25) Limonene, D (R); (26) Pyridine-d5; (27) Pyridazine; (28) N-Methylaniline; (29) N,N,N',N'-Tetramethylethylenediamine (TEMED); (30) 1,4-Dioxane; (31) n-Butylamine.

Table 8: Inliers found by sequential RANSAC in Hierarchical clusters 1 and 4. Each row lists the cluster, RANSAC iteration, inlier index and the corresponding compound with its feature values.

Cluster	Inlier number	Local index	Name
1	1	0	Acetonitrile-d3
1	2	4	Acetylacetone
1	3	8	Benzonitrile
1	4	10	Dimethylformamide (DMF)
1	5	2	N-Methyl-2-pyrrolidone (NMP)
1	6	9	Acetonitrile
1	7	11	Dimethyl sulfoxide (DMSO)
2	1	12	Water
2	2	14	Heavy water (D2O)
2	3	15	Heavy water (D2O)
2	4	17	Hydrogen peroxide
3	1	21	Pyridine
3	2	24	Tetrahydrofuran (THF)
3	3	30	Benzene-d6 (Deuterated benzene)
3	4	31	Benzene
3	5	33	Phosphorus tribromide
3	6	35	Triethylamine (Et3N)
3	7	36	Triethylamine (Et3N)
3	8	40	Acetic acid
3	9	42	Ethyl acetate
3	10	18	Piperidine
3	11	19	Pyrrolidine
3	12	28	Dicyclohexylamine
3	13	38	Tri-n-butyl phosphate (TBP)
3	14	41	Limonene, D (R)
3	15	22	Pyridine-d5
3	16	25	Pyridazine
3	17	26	N-Methylaniline
3	18	32	N,N,N',N'-Tetramethylethylenediamine (TEMED)
3	19	27	1,4-Dioxane
3	20	34	n-Butylamine
3	21	37	1,4-Dioxane

4.3 Sequential RANSAC for DBSCAN

The first DBSCAN cluster was used as the last cluster passed through the RANSAC algorithm. The results are presented in Figure 17 and Table 9.

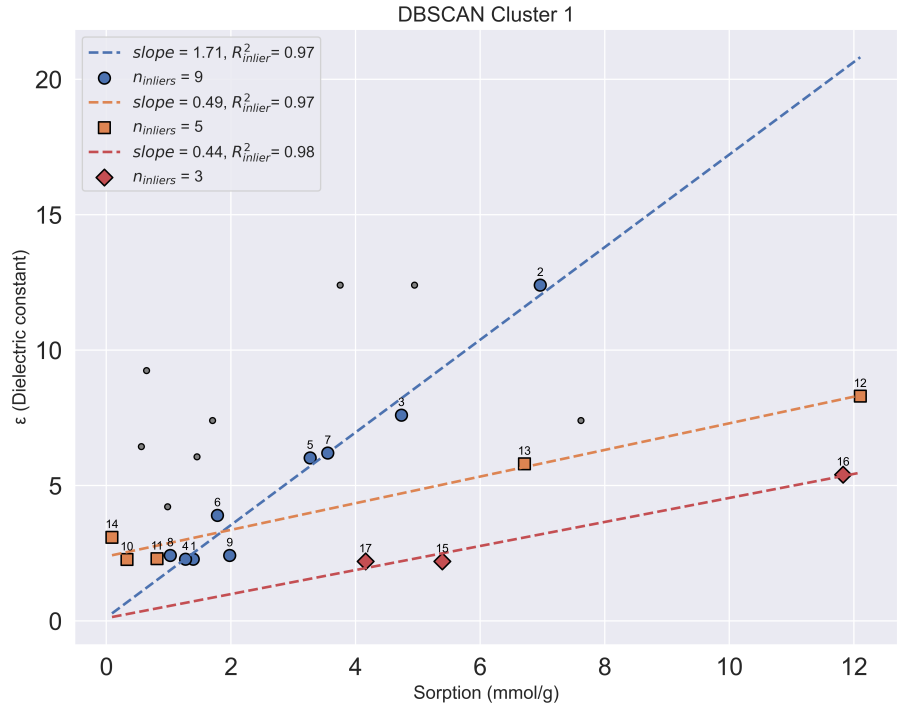


Figure 17: Sequential RANSAC Regression within the Hierarchical cluster 1.

Table 9: Sequential RANSAC Regression within the DBSCAN cluster 1. Each row lists the cluster, RANSAC iteration, inlier index and the corresponding compound with its feature values.

Cluster	Local index	Original index	Compound name
1	1	3	Benzene-d6 (Deuterated benzene)
1	2	12	Pyridine
1	3	15	Tetrahydrofuran (THF)
1	4	17	Benzene
1	5	18	Ethyl acetate
1	6	19	Phosphorus tribromide
1	7	20	Acetic acid
1	8	23	Triethylamine (Et_3N)
1	9	24	Triethylamine (Et_3N)
1	10	5	Dicyclohexylamine
1	11	6	Limonene, D (R)
1	12	14	Pyrrolidine
1	13	16	Piperidine
1	14	25	Tri-n-butyl phosphate (TBP)
1	15	9	1,4-Dioxane
1	16	22	n-Butylamine
1	17	27	1,4-Dioxane

The marked compounds in the figure are, in order: (1) Benzene-d6 (Deuterated benzene); (2) Pyridine; (3) Tetrahydrofuran (THF); (4) Benzene; (5) Ethyl acetate; (6) Phosphorus tribromide; (7) Acetic acid; (8) Triethylamine (Et_3N); (9) Triethylamine (Et_3N); (10) Dicyclohexylamine; (11) Limonene, D (R); (12) Pyrrolidine; (13) Piperidine; (14) Tri-n-butyl phosphate (TBP); (15) 1,4-Dioxane; (16) n-Butylamine; (17) 1,4-Dioxane.

5 Models Adjustment

Spectral clustering and DBSCAN are powerful clustering techniques suitable for data with unusual structure. In our work, of these two methods, only DBSCAN coped with the clustering task most satisfactorily, while spectral clustering showed unsatisfactory results. To improve the operation of this model, the data was standardized and new values of the adjacency matrix were calculated. Figures 18 and 19 are provided for StandartScaler and RobustScaler.

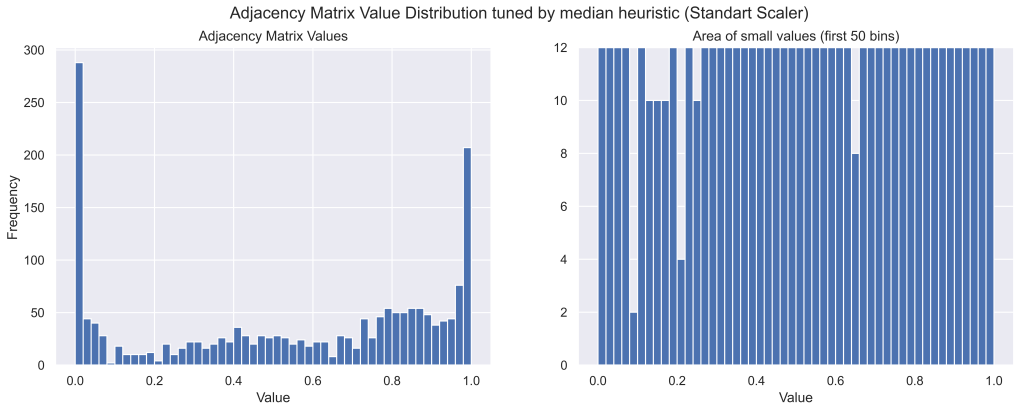


Figure 18: Adjacency Matrix Value distribution (StandartScaler).

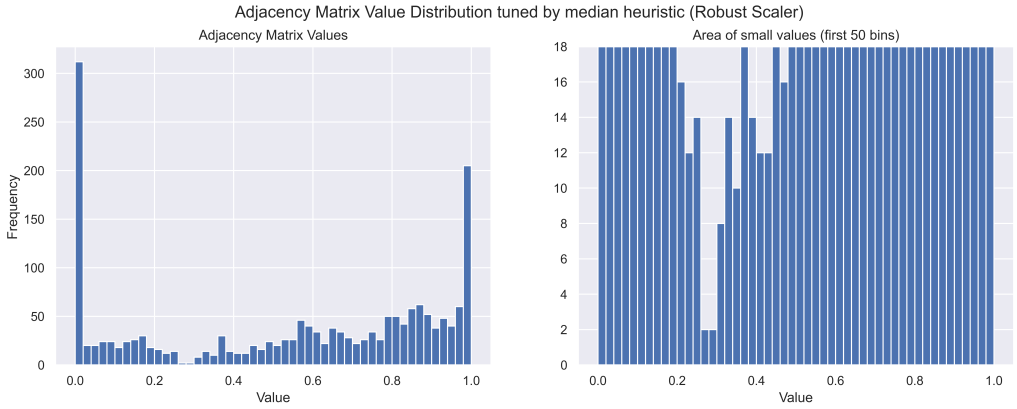


Figure 19: Adjacency Matrix Value distribution (RobustScaler).

Despite the improved calculations, the values of the adjacency matrices are mostly in zero and one, which indicates further anomalous eigenvalues of the Laplacian for us. There has not been much improvement in metrics for DBSCAN, since it is better to select parameters manually based on the meaning of the data. In any case, the bottom line for us is that both models don't handle the noisy data that is ours well.

References

[1] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622:178–210, 2023.

[2] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[3] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, volume 14, pages 849–856, 2002.

[4] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 226–231. AAAI Press, 1996.

[5] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, June 1981.