

Article

A Geometric Clustering Method of Lattice Fringes for Quantitative Assessment of the Active Phase Morphology of Sulfide Catalysts from HRTEM Images

Oleg Seredin ¹, Andrey Kopylov ^{1,*}, Olesia Kushnir ¹, Mariia Kniazeva ², Mikhail Boyko ²
and Anton Maximov ²

¹ Laboratory of Cognitive Technologies and Simulation Systems, Tula State University, Lenin Ave. 92, Tula 300012, Russia; oseredin@yandex.ru (O.S.); kushnir-olesya@rambler.ru (O.K.)

² Laboratory of Hydrocarbon Chemistry, A. V. Topchiev Institute of Petrochemical Synthesis, Russian Academy of Sciences, Leninsky Ave. 29, Moscow 119991, Russia; knyazeva@ips.ac.ru (M.K.); boyko@ips.ac.ru (M.B.); max@ips.ac.ru (A.M.)

* Correspondence: and.kopylov@gmail.com

Abstract

This paper proposes a geometric clustering method for lattice fringes in HRTEM images to enable quantitative morphological analysis of the active phase of sulfide catalysts directly from micrographs. We developed a composite distance function between polyline representations of lattice fringes that incorporates structural and morphological characteristics of the catalyst and possesses a direct geometric interpretation. The proposed function satisfies semimetric axioms, ensuring its compatibility with density- and graph-based clustering algorithms that do not require centroid representations. For grouping of manually annotated lattice fringes, the DBSCAN algorithm was employed with a minimum cluster size of one. Experimental validation on both synthetic and real-world data, as well as an ablation study, confirms the high accuracy and physical interpretability of the approach, establishing it as a promising tool for high-throughput morphological analysis of sulfide catalysts and rational design of catalytic systems.

Keywords: density-based clustering; HRTEM image analysis; sulfide catalysts; nanoparticle morphology; relational clustering; interpretable machine learning

1. Introduction

Catalytic processing of biomass [1], lignin [2], and pyrolysis oils [3] as alternative feedstocks for commercial fuels and petrochemical intermediates is associated with the need to remove a significant amount of heteroatomic compounds (oxygen-, sulfur-, and nitrogen-containing). Cobalt- or nickel-promoted molybdenum or tungsten disulfide nanostructured catalysts are the most widely used industrial systems [4,5]. Based on current knowledge, the catalytic activity of these systems is determined by the morphology of the active phase particles, in particular the average sulfide slab length and the stacking number [6].

High-resolution transmission electron microscopy (HRTEM) is a primary technique for the morphological study of sulfide catalysts, allowing for the acquisition of images of the active phase of nanoparticles [7]. These images have clearly defined structural features (see Figure 1) corresponding to the physicochemical structure of the material. Namely, Molybdenum (MoS₂) and tungsten (WS₂) sulfides, along with their promoted forms (NiMoS, CoMoS, NiWS), exhibit a layered structure [8]. Each individual layer follows a “sandwich”



Academic Editor: Alessandro Lo Schiavo

Received: 14 May 2026

Revised: 30 June 2026

Accepted: 1 July 2026

Published: 14 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

configuration: a central plane of molybdenum atoms is sandwiched between two sulfur planes, forming S–Mo–S bonds. The threadlike lattice fringes in HRTEM micrographs effectively represent MoS₂/WS₂ layers whose basal planes are oriented parallel to the electron beam [9]. An isolated fringe or a group of parallel fringes corresponds to catalyst nanoplatelets with specific spatial periodicity and varying orientations.

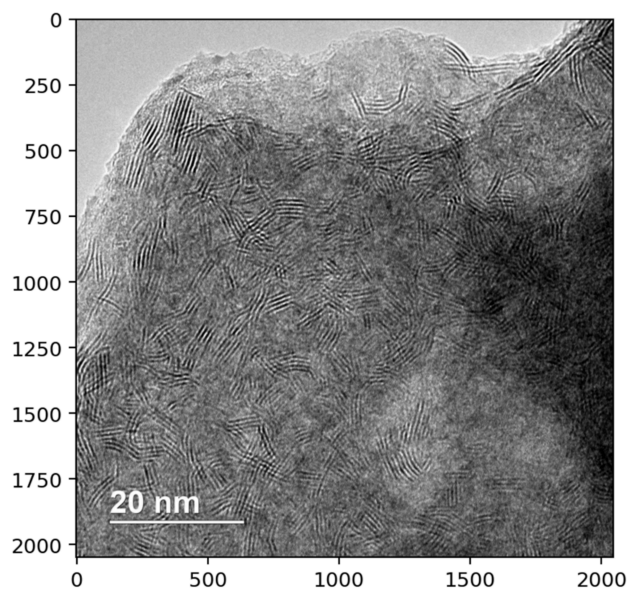


Figure 1. Example of an HRTEM image of a sulfide catalyst nanostructure.

In recent years, numerous automated and AI-assisted methods for HRTEM and STEM image analysis have appeared. However, almost all of them have been developed for other objects and address different tasks. For example, Maksov et al. [10] used deep learning to detect and classify defects in WS₂ and to follow phase evolution under the electron beam, but their focus is on defect dynamics rather than on particle morphology (slab length and stacking number). Larsen et al. [11] and Huang et al. [12] developed machine learning approaches for exit wave reconstruction and for determining slip/twist at van der Waals interfaces in MoS₂/WS₂. Other recent works target layer counting from optical [13], hyperspectral PL clustering [14], or electronic property prediction [15], and none of them address the problem of grouping fringes into sulfide stacks from HRTEM lattice fringes.

The closest related methodologies belong to the field of image-based nanocrystallography (IBN), which focuses on crystal structure identification and parameter determination through lattice fringe fingerprinting [16]. Zhong et al. [17] applied semantic segmentation (DeepLab V3+) to extract lattice fringes from HRTEM images of coal, but their work does not concern MoS₂/WS₂ catalysts nor the grouping of fringes into stacks. GRATEv2 [18] is an automated platform for analyzing conjugated polymers, where key segmentation features are fine fringes and their connectivity (Figure 2b). These studies primarily target crystallographic planes that appear as fringes due to phase contrast effects, yielding visual similarities to layered sulfide nanoparticles (Figure 2a). However, directly applying these methods to sulfide catalyst analysis is not feasible. In contrast, Ni(Co)Mo(W)S catalysts appear in micrographs as layered nanoplatelets with characteristic parallel fringes, often containing defects, curvatures, and diverse stacking structures (Figure 2c). Their analysis therefore requires an approach specifically designed for grouping fringes into stacks (sulfide stacks) and for extracting slab length and stacking number.

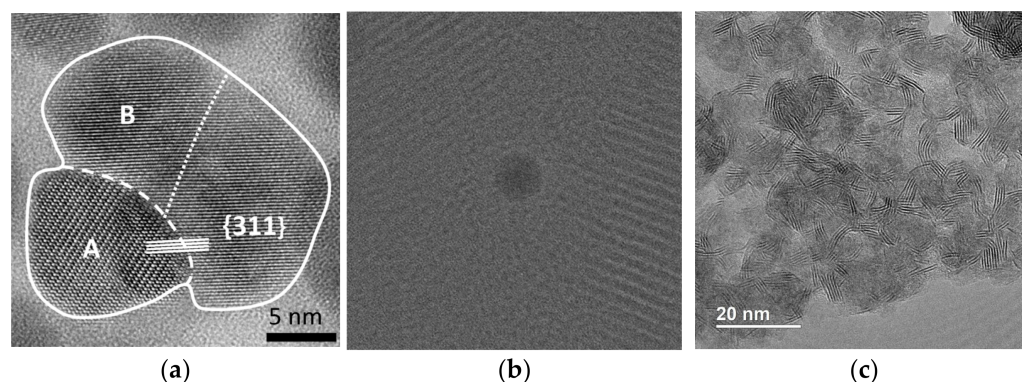


Figure 2. (a) HRTEM image of an iron oxide nanoparticle that consists of two cores, A and B from [19]; (b) HRTEM image of a conjugated polymer copolymer from [18]; (c) HRTEM image of in situ prepared NiMoS_{0.5} catalysts.

Thus, these studies demonstrate the rapid development of automated and AI-assisted microscopy data analysis. However, they predominantly address defect recognition, phase identification, crystallographic characterization, image reconstruction, or generic segmentation tasks.

In contrast, the objective of the present work is the grouping of lattice fringes into stacks and the extraction of catalyst-relevant morphological descriptors—namely slab length and stacking number. To the best of our knowledge, no previous study has proposed a physically motivated clustering framework specifically designed for this task.

Manual analysis of HRTEM images is highly labor-intensive and tedious. Consequently, the automation of the detection and analysis of TEM images for sulfide catalysts, specifically (i) detecting individual fringes corresponding to sulfide layers and (ii) aggregating them into groups corresponding to sulfide stacks, is a pressing need.

This work addresses one essential step in automated HRTEM image analysis, namely task (ii): merging layers represented by parallel fringes into groups corresponding to multilayered slabs. To solve this, we propose a specially designed composite distance function between polylines corresponding to detected layers. For MoS₂ and WS₂, the spatial arrangement of lattice fringes in HRTEM images (distance, parallelism, length ratios, and lateral shifts) obeys a set of physicochemical constraints that must be considered when constructing a distance function.

First of all, the interlayer spacing is fixed and lies in the range of 0.62–0.65 nm, as confirmed by HRTEM measurements (0.63 nm—Peng et al. [20]; 0.616–0.630 nm—Wang et al. [21]; see also Oviroh et al. [22]). The catalytic activity of sulfides is directly related to the average slab length and the stacking number [23,24], which implies that within a single nanoparticle, the layers have comparable lengths and are predominantly parallel. At the same time, due to stacking faults, the layers are rarely perfectly aligned; the works of Ramos et al. [25] and Isshiki et al. [26,27] provide evidence of shifts between S/S layers and the presence of edge dislocations. The ability of the MoS₂ lattice to bend [28] further justifies the need for a metric that is robust to small deviations from strict parallelism and that accounts for lateral shifts. All the above constraints are encoded in a specialized semimetric distance function between polylines.

Notably, using a distance function instead of explicit features (length, inclination, curvature, etc.) shifts the paradigm from traditional feature-based analysis to a featureless or relational-based approach [29], relying solely on pairwise object relationships. In this framework, merging polylines into particles becomes a metric clustering problem [30].

The semimetric properties of the proposed distance function make it impracticable to apply clustering methods based on Euclidean geometry and a predefined number of

clusters. Among methods supporting featureless representations and requiring no prior specification of the number of clusters (corresponding to nanoparticle count), DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [31] is widely used. DBSCAN relies on local point density, grouping densely packed points while identifying outliers.

Experimental results demonstrate strong agreement between expert annotations and the clustering results obtained with the proposed algorithm.

2. Task Formulation: Automated HRTEM Image Analysis for Sulfide Catalysts

Understanding the relationship between the nanostructure of the NiMoS phase and its activity enables the targeted design of highly active unsupported catalysts. HRTEM images of such particles exhibit a characteristic appearance with clearly defined parallel fringes corresponding to catalyst nanoplatelets with fixed spatial periodicity and varying orientations (see Figure 3).

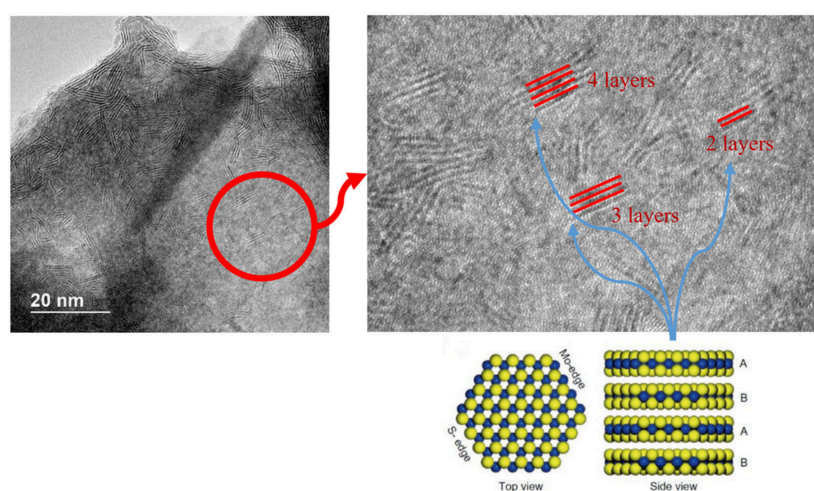


Figure 3. Structure of a TEM image of a sulfide catalyst and MoS₂ crystallite in a ball model with top and side views of a slab consisting of four layers (Adapted from [32]). Yellow balls represent sulfur atoms; blue balls represent molybdenum atoms.

The distribution of sulfide nanocrystallites by slab length L_m and stacking number N_m is a key characteristic of sulfide catalysts, as it forms the distribution of active site types on their surface and directly determines catalytic performance [23,24]. Catalysts predominantly featuring fewer layers often exhibit higher activity than those containing multilayered particles [33,34]. Furthermore, curved platelets frequently contain defects (dislocations, basal plane ruptures), which increase the number of accessible active sites and influence activity [35].

Analysis results of HRTEM micrographs are traditionally presented as histograms of detected nanoparticles by length and stacking number (Figure 4).

Given the large number of target objects in a micrograph and the ambiguity of human perception of such structures, manual annotation is extremely laborious. Automation requires solving at least two distinct problems: (i) detecting individual lines corresponding to layers, and (ii) merging detected lines into structures corresponding to sulfide nanoparticles. Two primary strategies naturally emerge. The first involves segmenting image regions corresponding to individual particles, followed by line detection within these regions, thereby transforming the second problem into region segmentation. The second strategy reverses this order: individual lines (layers) are first detected across the entire image, then merged into multilayered particles. Unmerged lines correspond to single-layer

particles. The second strategy is preferable due to the relative simplicity of handling object overlapping and intersection in electron microscopy images.

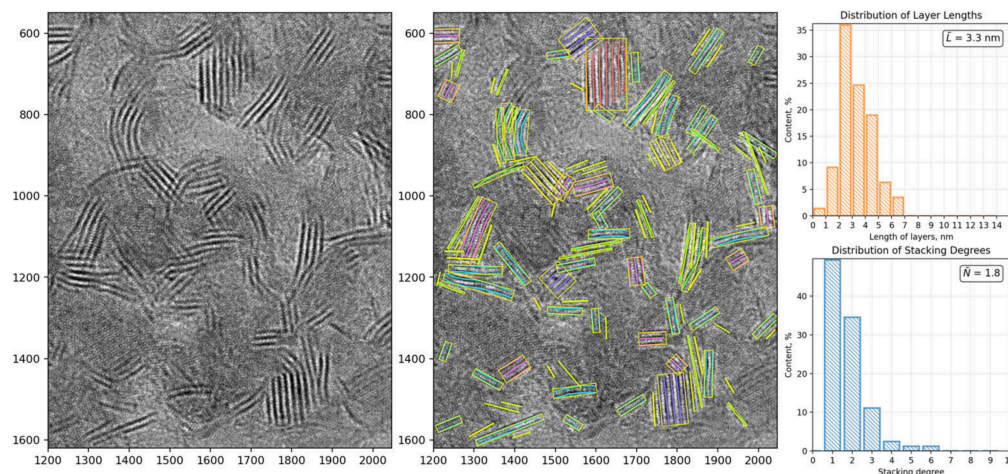


Figure 4. Manually annotated image fragment and histograms of layer length and stacking number per particle. Different line colors correspond to the number of layers in particles (highlighted by yellow rectangles).

This work adopts the second strategy and focuses on the second step: merging pre-detected lines into catalyst particle structures. From a machine learning perspective, this is framed as a clustering problem. Objects in such problems can be represented either as fixed feature vectors or via pairwise relationships. We represent the projection of a nano stack slab onto the image plane as a polyline l with ordered vertices (a vertex tuple) defined by coordinates (x_i, y_i) , $i = 1, 2, \dots, n$. Redundant collinear vertices between two neighbors are removed via a simple preprocessing check. Duplicate or reversely ordered vertex tuples are also excluded. This representation is optimal for both expert manual annotation and computational storage/analysis.

While this representation can serve as input for standard clustering algorithms, the nature of the object and HRTEM visualization impose constraints on how nanoplatelets should be merged:

1. Interlayer distance is fixed for a given catalyst composition, e.g., the Inorganic Crystal Structures Dataset (ICSD) [36] lists the MoS_2 lattice parameter $c \approx 12.3201312.9 \text{ \AA}$, yielding an interlayer spacing of 0.615–0.645 nm, other studies reported similar values (0.63 nm—Peng et al. [20]; 0.616–0.630 nm—Wang et al. [21]; see also Oviroh et al. [22]).
2. Layers may be curved, not strictly straight. Hilton and Fleischauer [28] showed that the MoS_2 lattice “can bend, kink, or twist”.
3. Layers are arranged parallel to each other.
4. Layer lengths within a particle are uniform. Wang et al. [23] explicitly state that the catalytic activity of MoS_2 is determined by the average slab length and the stacking number. To compute an average, the layers within a given particle must have similar lengths and be predominantly parallel—otherwise the concept of “average length” for a nanoparticle loses its meaning. This is a widely accepted notion in modern literature on sulfide catalyst properties.
5. Lateral shifts between layers are rare. The works of Ramos et al. [25] and Isshiki et al. [26,27] provide evidence of shifts between S/S layers and the presence of edge dislocations.

These five assumptions are reliably supported by experimental and literature evidence and reflect the actual structural characteristics of MoS_2 and WS_2 . Encoding these constraints

directly into a vertex-coordinate feature space is complex and requires modifying standard clustering algorithms. Additionally, vertex numbering order may vary across lines. A featureless approach based on pairwise comparison treats each polyline as an independent object regardless of image coordinates, while embedding the constraints into a custom two-argument function acting as a distance in the relational space.

3. Distance Function Between Polylines

Given these considerations, we define the distance $D(l_i, l_j)$ between two layers l_i and l_j , $i, j = 1, 2, \dots, L$, where L —is the total number of detected layers as a function composed of multiplicative components, each reflecting a specific nanoparticle structural feature:

$$D(l_i, l_j) = d_{gap}(l_i, l_j) \cdot R_{len}(l_i, l_j) \cdot R_{dist}(l_i, l_j) \cdot R_{diag}(l_i, l_j), i, j = 1, 2, \dots, L. \quad (1)$$

Let us examine the function structure in more detail.

$d_{gap}(l_i, l_j)$ is the base component of the distance function dependent on the distance between the polylines. The term “distance between polylines” requires clarification. Since different polylines can be represented by different numbers of vertices and have different lengths, the task arises of finding points of polylines corresponding to each other, or, otherwise, the alignment task. To avoid computationally expensive dynamic alignment, we propose to divide each polyline into an equal number of equal-length intervals and measure the average Euclidean distance between the points of the polylines lying on the boundaries of the intervals with matching indices. The coordinates of these points are determined by linear interpolation of the coordinates of the corresponding neighboring vertices of the polyline. To resolve vertex ordering ambiguity, the average is calculated twice—for the forward and reverse orders—and the minimum of them is selected as the final estimate.

Let P be the number of vertices of a polyline. These vertices divide a polyline into $P - 1$ intervals. Denote $(x_{l_1,k}, y_{l_1,k})$, $k = 1, 2, \dots, P$ the coordinates of the vertices of the polyline l_1 , and $(x_{l_2,k}, y_{l_2,k})$, $k = 1, 2, \dots, P$ for the polyline l_2 . Therefore $d_{i,j}(l_1, l_2) = \sqrt{(x_{l_1,i} - x_{l_2,j})^2 + (y_{l_1,i} - y_{l_2,j})^2}$ is the Euclidean distance between points i and j . Then the mean distance between these lines d_{mean} can be defined as:

$$d_{mean}(l_1, l_2) = \frac{\min\left(\sum_{k=1}^P d_{k,k}(l_1, l_2), \sum_{k=1}^P d_{k,P-k+1}(l_1, l_2)\right)}{P}. \quad (2)$$

The numerator in Formula (2) determines the order of numbering points of compared polylines. We assume that after the calculation of (2), this order is determined and used when calculating the remaining components of function $D(l_1, l_2)$ (1). In this case, we can, where appropriate, write $d_k(l_1, l_2)$ instead of $d_{k,k}(l_1, l_2)$.

If the mean distance between polylines significantly exceeds the standard reference value, one can avoid further calculations and assign the distance function $D(l_1, l_2)$ a sufficiently large value, for example, based on the dimensions of the analyzed image, as a rough estimate of the distance between the most distant polylines.

We normalize the $d_{mean}(l_1, l_2)$ value by the interlayer distance w in the nanoparticle to obtain values close to unity regardless of its type. Since the interlayer distance in a particle is subject to constraints, we significantly increase the weight of this component in the final distance function using an exponential dependence on $d_{mean}(l_1, l_2)$. As a result, the component responsible for the distance between polylines takes the form:

$$d_{gap}(l_i, l_j) = e^{d_{mean}(l_i, l_j)/w} - 1.$$

The remaining components of the function (1) serve as correcting coefficients that allow for accounting for features of topological arrangement and nanoparticle morphology, so they are selected such that their minimum value equals 1.

$R_{\text{len}}(l_i, l_j)$ penalizes combining layers that differ significantly in their length. This component represents the ratio of the maximum and minimum lengths of the compared polylines.

Denote $s(l)$ as the length of the line l , defined as the sum of Euclidean distances between its neighboring vertices, then:

$$R_{\text{len}}(l_i, l_j) = \frac{\max(s(l_i), s(l_j))}{\min(s(l_i), s(l_j))}.$$

The mean distance between polylines does not allow us to fully determine the specificity of their mutual arrangement. The requirement of layer unidirectionality, considering that lines can be curved, can be expressed by the following component reflecting the homogeneity of distances:

$$R_{\text{dist}}(l_i, l_j) = \frac{\max_{k=1, \dots, P} (d_k(l_1, l_2)) + \xi}{\min_{k=1, \dots, P} (d_k(l_1, l_2)) + \xi}.$$

A small regularizing constant $\xi > 0$ serves to prevent division by zero while preserving the minimum component value equal to 1.

The next component, $R_{\text{diag}}(l_i, l_j)$, penalizes shifts in lines relative to each other while preserving distance homogeneity and is defined as the ratio of “diagonal” lengths, that is, the distance between the initial point of the first line and the final point of the second and the final point of the first line and the initial point of the second line, considering the direction of numbering found in (2):

$$R_{\text{diag}}(l_i, l_j) = \frac{\max(d_{1,P}(l_i, l_j), d_{P,1}(l_i, l_j))}{\min(d_{1,P}(l_i, l_j), d_{P,1}(l_i, l_j))}.$$

Obviously, if the “diagonals” do not intersect, the lines cannot belong to the same particle due to nanoparticle structure features, and the overall distance estimate $D(l_1, l_2)$ should be assigned a sufficiently large value, for example, based on the dimensions of the analyzed image.

The characteristics of distance function components are illustrated in Figure 5.

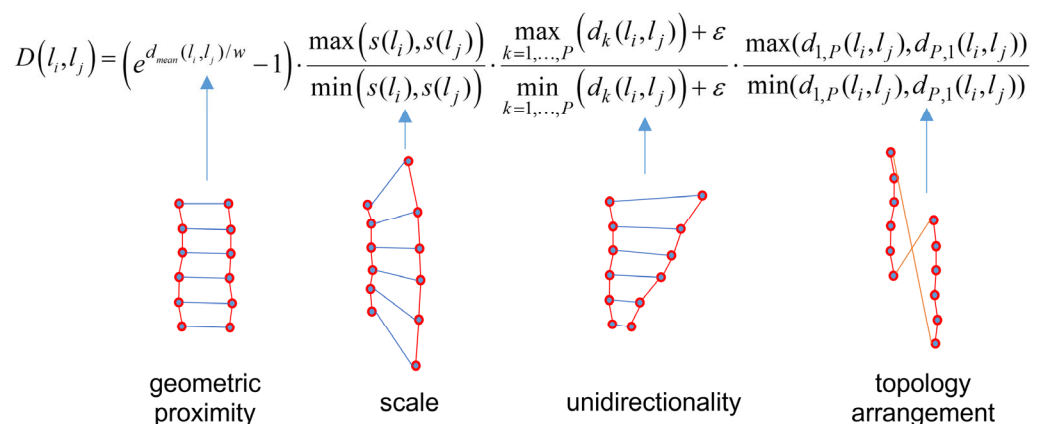


Figure 5. Characteristics of distance function components.

The proposed distance function $D(l_i, l_j)$ satisfies the following properties:

- Non-negativity: $D(l_i, l_j) \geq 0$ for any polylines l_i, l_j , since all components of the proposed function take values in $[1, \infty)$ or $[0, \infty)$;
- Identity: $D(l_i, l_j) = 0$ if and only if l_i and l_j coincide with accuracy up to the direction of traversal, since in this case $d_{gap}(l_i, l_j) = 0$, and all other components equal unity;
- Symmetry: $D(l_i, l_j) = D(l_j, l_i)$, since all components are symmetric with respect to argument permutation;
- Triangle inequality: not guaranteed due to the multiplicative-exponential structure of the function.

Thus, $D(l_i, l_j)$ is a semimetric, which imposes certain limitations on the type of applicable clustering algorithms.

4. Polylines Clustering

Modern clustering methods can be divided into the following groups, depending on how they interpret the compactness hypothesis [37]:

- Graph theory-based approaches: Clusters are defined as connected components or as complete subgraphs (i.e., cliques) of corresponding graphs;
- Centroid approaches: A cluster as some “averaged vector” in the corresponding linear space;
- Density-based approaches: A cluster is a family of regions with a certain point density in the corresponding linear space;
- Distribution-based approaches: A cluster is modeled using a multidimensional statistical distribution (in the overwhelming majority of cases, the normal distribution) in the corresponding linear space.

Due to the features of the distance function, the proposed distance function forms a semimetric space, and clustering algorithms that are oriented toward centroid approaches and minimization of intra-class dispersion, such as K-means [38], Ward’s method [39], or the EM algorithm [40,41], cannot be successfully applied, since they become mathematically ill-defined in such spaces for the following reasons:

K-Means relies on computing centroids (arithmetic means) to minimize within-cluster variance. In a semimetric space, the triangle inequality does not hold, meaning the arithmetic mean no longer represents the geometric center that minimizes the sum of distances. Furthermore, it requires pre-defining the number of clusters, which is unacceptable for the problem under consideration.

Standard EM-based clustering assumes data is generated from multivariate Gaussian distributions. The covariance matrices and the associated Mahalanobis distance strictly require a metric space derived from an inner product. Consequently, the probabilistic density functions are ill-defined when the distance function lacks strict metric properties. While theoretically possible, Kernel-based EM is generally ill-suited for this specific case because the distance function likely does not satisfy the conditions required to form a valid kernel.

Ward’s method is designed to minimize the total within-cluster variance (the sum of squared errors). The distance update formula for Ward’s method is derived directly from Euclidean geometry and relies on the existence of a well-defined geometric centroid that minimizes squared distances, which is not the case for semimetric spaces.

Thus, for the problem being solved, both graph theory-based algorithms and density-based algorithms can be chosen. These methods rely solely on pairwise distance comparisons to define ϵ -neighborhoods and reachability. They do not compute centroids, assume global probability distributions, or require hierarchical monotonicity, making them inherently robust and mathematically appropriate for clustering in semimetric spaces.

There is a deep theoretical connection between graph-based and density-based clustering algorithms [42], and in many ways, they can be considered as one class of methods.

Thus, the DBSCAN algorithm [31] with a minimum number of objects required for cluster formation equal to one is completely equivalent to hierarchical clustering with single linkage, which is implemented through breaking edges of the minimum spanning tree. Since in the task of analyzing HRTEM images of sulfide catalysts, single-layer particles are present, a setting with a minimum number of elements in a cluster equal to one is allowed, unlike those tasks where single objects are considered noise.

As a result of the clustering procedure, we obtain a partition of a set $\mathcal{L}$ of all polylines into subsets $\{C_t\}_{t \in \{1, \dots, T\}}$, where T is the number of clusters found. $\cup_{t \in \{1, \dots, T\}} C_t = \mathcal{L}$, $C_t \cap C_s = \emptyset, \forall t \neq s$, $C_t \neq \emptyset, \forall t$. Each subset corresponds to a nanoparticle slab, and its cardinality represents the stacking number.

5. Overview of the Proposed Framework

The proposed framework is described in Algorithm 1. The process begins by resampling each polyline to a uniform number of vertices via linear interpolation to ensure consistent representation. A symmetric distance matrix is then computed for all pairs of polylines using a proposed specific distance metric (1). This matrix serves as the input for the DBSCAN clustering algorithm, which partitions the set of polylines into disjoint subsets corresponding to individual nanoparticle slabs. Finally, the stacking number for each particle is determined by calculating the cardinality of its respective cluster.

Finally, the morphological descriptors—such as average slab length and mean stacking number—are aggregated across all identified clusters to provide a comprehensive quantitative characterization of the catalyst's active phase. These statistics can then be used for comparative analysis between different catalyst samples or synthesis conditions.

Algorithm 1. Proposed framework for quantitative assessment of the active phase morphology of sulfide catalysts

Input: A set of polylines $\mathcal{L} = \{l_i = (x_{l_i,k}, y_{l_i,k}), k = 1, 2, \dots, P_i, i = 1, 2, \dots, L\}$, where L —number of polylines,

P_i —number of vertices in the polyline l_i .

Output: 1. A partition of a set $\mathcal{L}$ into subsets $\{C_t\}_{t \in \{1, \dots, T\}}$, where T is the number of clusters (nanoparticle slabs) found. $\cup_{t \in \{1, \dots, T\}} C_t = \mathcal{L}$, $C_t \cap C_s = \emptyset, \forall t \neq s$, $C_t \neq \emptyset, \forall t$.

2. Stacking numbers N_t of all particles (equivalent to cluster cardinality).

1. Resample each polyline $l \in \mathcal{L}$ to P vertices using linear interpolation.
 2. For each pair of polylines $l_i, l_j \in \mathcal{L}$ compute a symmetric distance matrix $D = D(l_i, l_j), j < i, i, j = 1, 2, \dots, L$ according to Equation (1).
 3. Apply DBSCAN clustering algorithm with distance matrix D to obtain partition of a set $\mathcal{L}$ into subsets $\{C_t\}_{t \in \{1, \dots, T\}}$ (nanoparticle slabs).
 4. Compute cluster cardinalities $N_t = |C_t|, \forall t \in \{1, \dots, T\}$.
-

6. Results on Synthetic Data Experiments

6.1. Data Model

To evaluate the impact of the relative spatial configuration of polylines corresponding to lattice fringes on the distance function and its components, we conducted five controlled experiments on synthetic data, systematically varying a single geometric parameter at a time while holding others constant. In accordance with the constraints imposed by the physicochemical features of the active phase structure of sulfide catalysts, described in

Section 2, we identified the following key parameters: inter-center distance, length ratio, mutual orientation, parallel shift, and uniform scaling.

A pair of straight polylines (l_1, l_2) with the following parameters is used as the initial data (Figure 6):

- $\mathbf{c} = (c_x, c_y) \in \mathbb{R}^2$ —offset vector from the center of l_1 to center of l_2 ;
- $s(l_1), s(l_2) \in \mathbb{R}^+$ —the lengths of the first and second polylines;
- $\theta \in [-\pi, \pi]$ —orientation angle of l_2 relative to l_1 .

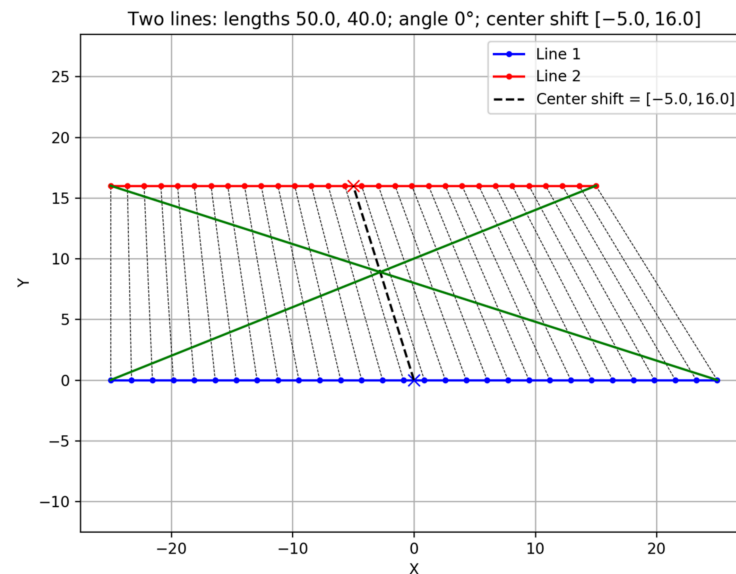


Figure 6. Example of synthetic lines configuration.

The first polyline is centered at the origin and oriented along the axis x :

$$l_1 = \left\{ \left(-\frac{s(l_1)}{2}, 0 \right), \left(\frac{s(l_1)}{2}, 0 \right) \right\}.$$

The second polyline is centered at a point $\mathbf{c}$ and is oriented at an angle θ :

$$l_2 = \left\{ \mathbf{c} - \frac{s(l_2)}{2} \cdot \mathbf{u}(\theta), \mathbf{c} + \frac{s(l_2)}{2} \cdot \mathbf{u}(\theta) \right\}, \mathbf{u}(\theta) = (\cos \theta, \sin \theta).$$

6.2. Experiment 1: Dependence on Inter-Center Distance

The experiment is aimed at verifying the fundamental property of the metric—the monotonic increase in distance with increasing spatial separation of lines. This property is critically important for clustering tasks, as it ensures the separation of objects located in different areas of space.

Vertical offset d of lines centers $c_y = d$ varied from 1 to 32 pixels (step 0.31). All other parameters were fixed: the lengths of both lines were $s(l_1) = s(l_2) = 20$ pixels, angle between lines $\theta = 0$ (parallel orientation), horizontal offset $c_x = 0$.

The dependence of the components of the distance function on the vertical shift in the line centers is shown in Figure 7.

For parallel lines of the same length, the components $R_{\text{len}}(l_i, l_j)$ (length ratio) and $R_{\text{diag}}(l_i, l_j)$ (the ratio of diagonals) remain close to unity, since the geometry of the configuration is symmetrical. The $R_{\text{dist}}(l_i, l_j)$ component (uniformity of distances) also varies slightly, since the pairwise distances between the sampled points are evenly distributed. The main factor determining the growth of the metric is the component $d_{\text{gap}}(l_i, l_j)$ which depends exponentially on the average distance.

The dependence graph demonstrates a monotonic exponential growth, which confirms the correctness of the chosen form of normalization of the average state. The rapid increase in metrics when lines are moved away from each other contributes to the formation of compact clusters in algorithms based on density or graph properties.

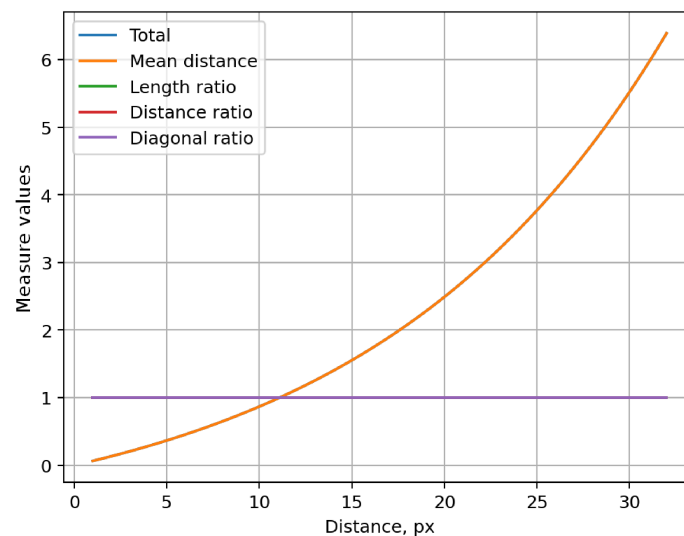


Figure 7. The dependence of the components of the distance function on the vertical shift in the line centers. Total measure coincides with mean distance component $d_{mean}(l_1, l_2)$. Length $R_{len}(l_1, l_2)$, distance $R_{dist}(l_1, l_2)$ and diagonal ratio $R_{diag}(l_1, l_2)$ are constant and equal to 1.

6.3. Experiment 2: Dependence on Length Ratio

The experiment evaluates the sensitivity of the metric to differences in line lengths. The length of the first line is fixed: $s(l_1) = 20$ pixels. The length of the second line $s(l_2)$ varies from 1 to 77 pixels in increments of 0.77. The remaining parameters are the vertical offset of the pixel centers $c_y = 16$, the horizontal offset $c_x = 0$, and the angle $\theta = 0$.

The component $R_{len}(l_1, l_2)$ (length ratio) in Figure 8 demonstrates a hyperbolic dependence:

$$R_{len}(l_1, l_2) = \begin{cases} \frac{20}{s(l_2)}, & s(l_2) < 20 \\ \frac{s(l_2)}{20}, & s(l_2) \geq 20 \end{cases}.$$

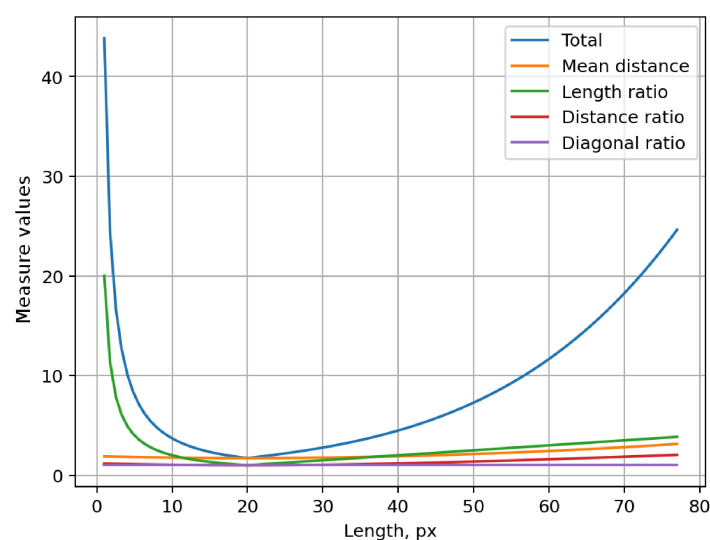


Figure 8. Distance function component dependence on second polyline length.

At $s(l_2) \rightarrow 20$, the ratio tends to unity, providing a minimal contribution to the metric. When deflected $s(l_2)$ in any direction, $R_{\text{len}}(l_1, l_2)$ grows linearly relative to the logarithm of the ratio of lengths. The component $R_{\text{dist}}(l_1, l_2)$ also increases with a strong difference in lengths, as the pairwise distances between the points become heterogeneous: the points at the ends of the longer line are further away from the corresponding points of the shorter line.

6.4. Experiment 3: Dependence on Orientation

The experiment investigates the response of the distance function to the angular divergence of lines. Orientation sensitivity is important to control the consistency of line directions when grouping.

Lengths of both lines are fixed: $s(l_1) = s(l_2) = 20$ pixels. The angle θ varies in the range from -30° to $+30^\circ$ (from $-\pi/6$ to $\pi/6$ radian) with step of 0.6° . Vertical centers offset $c_y = 16$ pixels, horizontal shift $c_x = 0$.

At $\theta = 0$, the lines are parallel, and the pairwise distances between the sampled points are uniform, which minimizes the $R_{\text{dist}}(l_1, l_2)$ and $R_{\text{diag}}(l_1, l_2)$ components. As the absolute value of the $|\theta|$ angle increases, the lines diverge, which leads to an inhomogeneity of the pairwise distances (growth of $R_{\text{dist}}(l_1, l_2)$) and the asymmetry of the diagonals (growth of $R_{\text{diag}}(l_1, l_2)$).

The graph in Figure 9 shows a symmetrical relationship with exponential growth acceleration at the edges of the range, with a minimum at $\theta = 0$. The symmetry with respect to zero reflects the fact that a turn in either direction is equivalent in terms of geometric similarity. The shape of the curve ensures a smooth increase in the “penalty” for deviation from parallelism, which contributes to stable clustering by orientation.

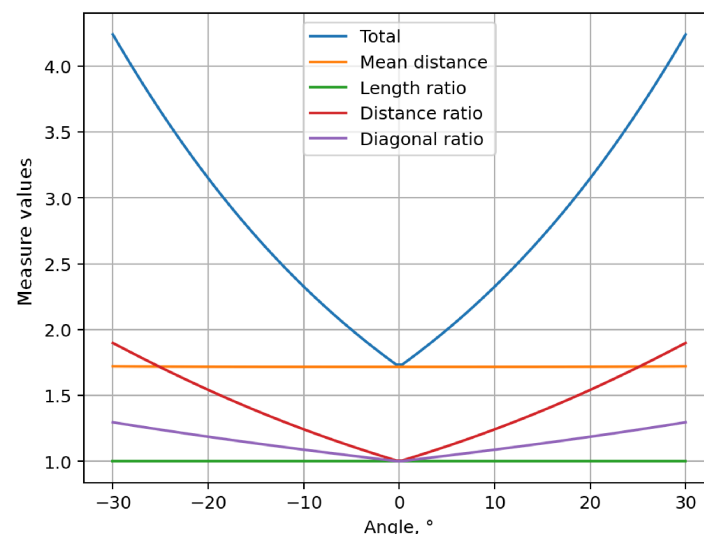


Figure 9. Distance function component dependence on the line’s orientation angle.

6.5. Experiment 4: Dependence on Parallel Shift

The experiment evaluates the effect of horizontal displacement of lines at a fixed vertical distance. This allows us to understand how the metric is able to distinguish parallel lines that are offset along its direction.

The lengths of both lines have been reduced to $s(l_1) = s(l_2) = 10$ pixels to enhance the shift effect. The horizontal offset $c_x = h$ ranges from -30 to $+30$ pixels. The vertical offset is fixed: $c_y = 16$ pixels. Angle $\theta = 0$ (parallel lines).

At $h = 0$, the lines are centered relative to each other, which minimizes the average distance and ensures the symmetry of the diagonals ($R_{\text{diag}}(l_1, l_2) \approx 1$). As the absolute

value of the shift $|h|$ increases, the points of one line shift relative to the corresponding points of the other line, which leads to an increase in the average distance and, accordingly $d_{gap}(l_1, l_2)$. The $R_{diag}(l_1, l_2)$ component also increases as one of the diagonals ($d_{1,p}(l_1, l_2)$ or $d_{p,1}(l_1, l_2)$) lengthens and the other shortens. The component $R_{dist}(l_1, l_2)$ grows due to the heterogeneity of the paired distances: the extreme points diverge more than the central ones.

The dependence graph in Figure 10 is symmetrical with respect to the zero offset and demonstrates monotonous growth as it increases.

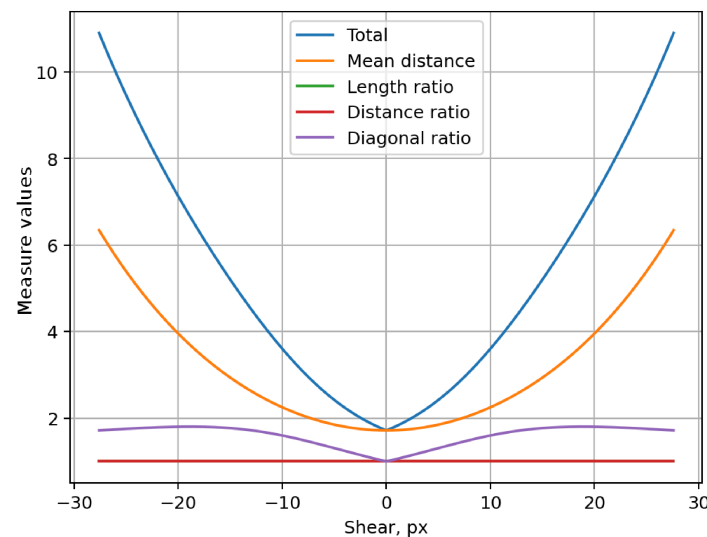


Figure 10. Distance function component dependence on parallel shift.

At $c_y = 16$ pixels, the growth is accelerated due to the $c_x = (s(l_2, \Delta) - s(l_1, \Delta))/2 = 5$ component, which reflects a qualitative change in the relative position of the lines: they cease to overlap in the projection to the direction. This property makes it possible to effectively distinguish between the offsets of parallel polylines.

6.6. Experiment 5: Dependence on Uniform Scaling

The experiment investigates the invariance of the function to a uniform scaling of the lengths of both lines. This property is important for applying the metric in problems where the absolute scale of objects may vary, but their relative geometry remains the same.

The line lengths are set as $s(l_1, \Delta) = 20 + \Delta$ and $s(l_2, \Delta) = 30 + \Delta$ where the parameter Δ varies from 0 to 300 pixels. The angle $\theta = 0$, the vertical displacement $c_y = 16$ pixels. The horizontal offset is selected as $c_x = (s(l_2, \Delta) - s(l_1, \Delta))/2 = 5$ pixels to compensate for the difference in line lengths and centering.

The behavior of the function in Figure 11 is due to the joint decrease of the $R_{len}(l_1, l_2)$ and $R_{diag}(l_1, l_2)$ components. With small Δ , the lines vary significantly in length (ratio 1.5), which makes a big contribution to the metric. With large Δ values, the absolute difference of 10 pixels becomes insignificant against the background of large lengths (~ 300 pixels), the ratio of lengths tends to 1, and the metric decreases.

This experiment demonstrates an important property of the proposed distance function, namely the predominant effect of the relative difference in lengths compared with the absolute one. Two lines with a length of 20 and 30 pixels are perceived by the metric as more different than lines with a length of 320 and 330 pixels, despite the fact that the absolute difference is the same (10 pixels).

This property is natural for many applications where the relative rather than absolute difference in the size of objects is important.

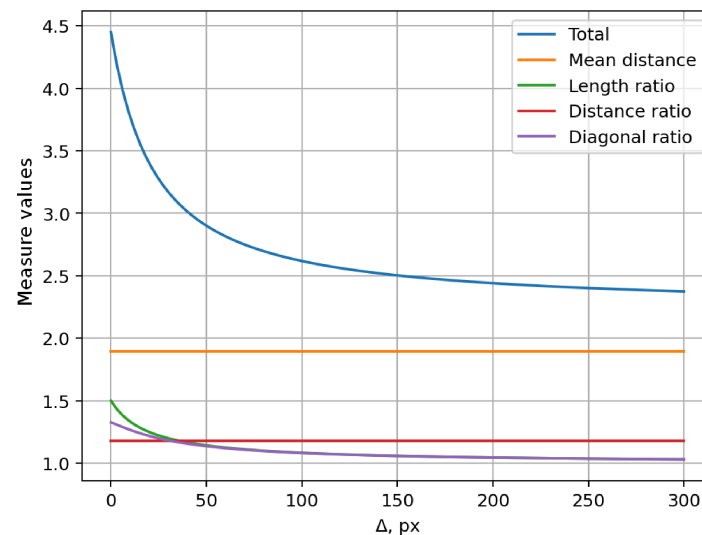


Figure 11. Distance function component dependence on uniform scaling.

6.7. Discussion on Experiments on a Synthetic Dataset

The results of five experiments allow us to draw the following conclusions about the sensitivity of the metric to geometric parameters:

1. The center-to-center distance has the strongest effect on the metric due to the exponential component $d_{gap}(l_1, l_2)$. This ensures effective separation of spatially distant lines.
2. The length ratio makes a linear contribution through the component $R_{len}(l_1, l_2)$. This makes it possible to group lines of similar length, but requires adjustment if the length is not a relevant feature.
3. The angular discrepancy causes the metric to grow due to the $R_{dist}(l_1, l_2)$ and $R_{diag}(l_1, l_2)$ components, which make the metric suitable for orientation clustering.
4. The parallel shift affects the metric symmetrically; the effect becomes greater when half the length of the line is exceeded, allowing us to distinguish the displaced nanoplates.
5. Uniform scaling has little effect on the metric while maintaining relative geometry, which ensures stability to variations in the absolute scale.

Thus, the distance function combines four geometrically interpreted components, providing monotonicity, discriminability, and relative scale invariance. Five model experiments systematically examine the sensitivity of the metric to key geometric parameters, confirming its adequacy for use in clustering algorithms such as DBSCAN and graph algorithms.

7. Results on Real Data Experiments

7.1. Dataset

Unfortunately, at the moment, there are no publicly available labeled data sets of high-resolution digital HRTEM images of molybdenum (or tungsten) sulfide catalysts that could be used for an objective assessment of the proposed layer clustering method.

Due to the high labor intensity of the process, scientific experts manually marked rectangular fragments of two images, C0_0023 and C0_0025, shown in Figure 12. Specifically, individual layers assembled directly into particles were marked. The marking was performed using the CVAT version 2.69.0 software tool (<https://www.cvat.ai/>, accessed on 14 February 2026), which contains a set of polylines corresponding to nanoplatelets, as well as the grouping of lines into particles (see Supplementary Materials). For ease of visual perception, layers belonging to particles of different stacking degrees (1, 2, 3, ...) are colored differently (see Figure 12).

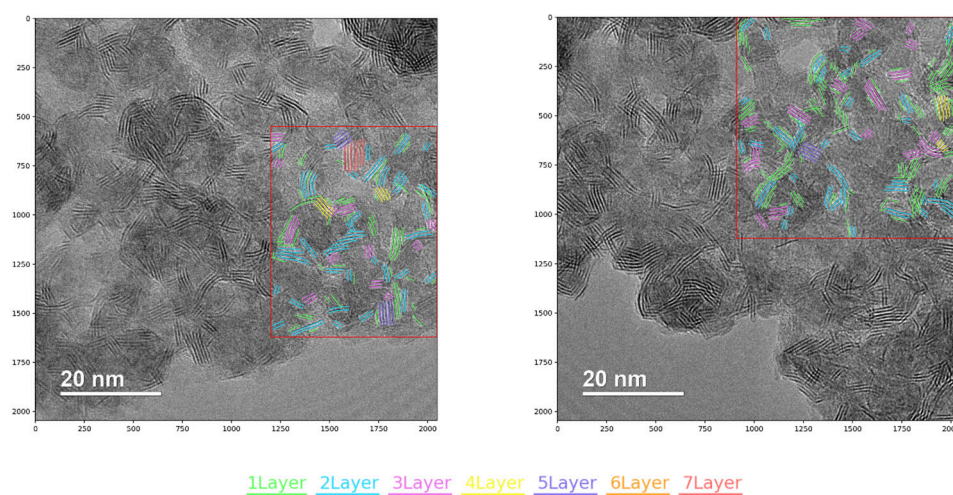


Figure 12. Manually annotated image fragments: C0_0023 (left), C0_0025 (right). The boundaries of the fragments are marked in red.

In image C0_0023, the expert found 273 polylines combined into 194 particles, and in image C0_0025, 196 polylines and 114 particles.

It is important to acknowledge that the manual annotations used as ground truth in this study are subject to inherent human uncertainty, particularly when interpreting ambiguous features such as overlapping fringes or low-contrast boundaries in HRTEM images. While the limited dataset size precluded a comprehensive statistical analysis of inter-expert variability, we mitigated potential labeling bias by implementing a rigorous double-checking protocol, wherein all annotated lines were independently verified by a second domain expert to ensure consistency and accuracy.

For each image, a matrix of distances between lines was formed according to expression (1). The standard distance between layers of catalyst particles of a given type and chemical composition, estimated from the image, was 16 pixels, which corresponds to the interlayer distance data in publicly available databases. [36]. The total number of vertices for all polylines after the resampling operation is set to 30, which corresponds to the minimum line length established by an expert. To automatically combine lines into particles, a clustering algorithm (DBSCAN) [31], in particular its scikit-learn Python (3.14.6) library implementation, was used with the minimum number of lines required to form a cluster equal to one and input data in the form of a distance matrix.

7.2. Calculation of Morphological Parameters in HRTEM Image Analysis

For each detected particle, two key morphological parameters are calculated:

- Slab length $\bar{s}$: the mean value of the lengths $s(l)$ for all layers l in the stack, scaled to nanometers (nm) using the image calibration.
- Stacking number N : number of layers in the particle, equivalent to cluster cardinality.

As a quality criterion for the proposed method, we will consider the similarity of the histogram of the distribution of particles by the number of layers obtained by manual marking by experts and obtained by our method (Figure 13).

To compare histograms, we apply two widely used metrics: the Kullback–Leibler divergence [43], which estimates the discrepancy between distributions from the information theory point of view, and the Vaserstein distance W_1 [44], also known as the earth mover’s distance (EMD) [45], which interprets the difference between a pair of histograms as the amount of work required to move the amount of “earth” so that the histograms become identical. The EMD value for a pair of identical histograms is zero. Figure 14 shows the de-

pendence of the histogram difference metrics on the value of ε the radius of a neighborhood of the DBSCAN algorithm.

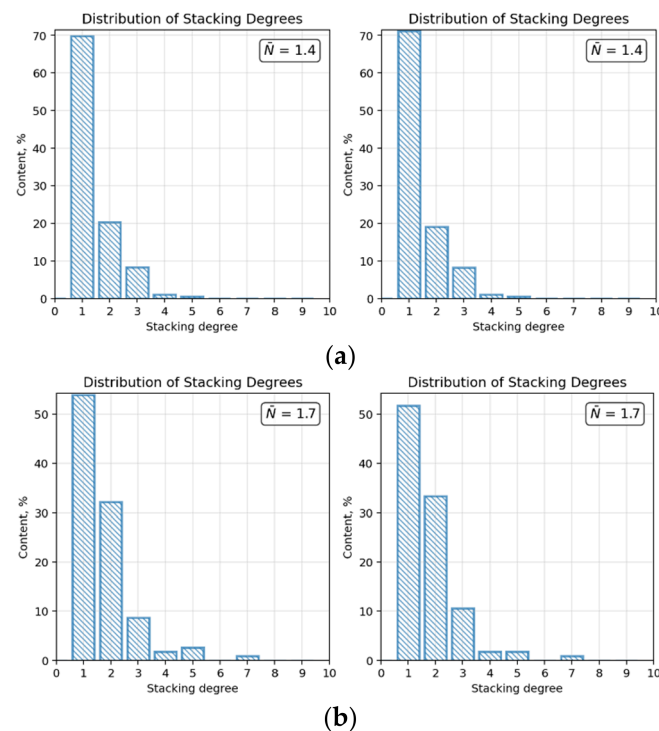


Figure 13. Histograms of the distribution of the number of layers in particles: (a)—for image C0_0023, (b)—for image C0_0025. The left column is formed by the DBSCAN algorithm with the $\varepsilon = 4.33$ parameter and the proposed metric (1), the right column is formed based on the expert's markup.

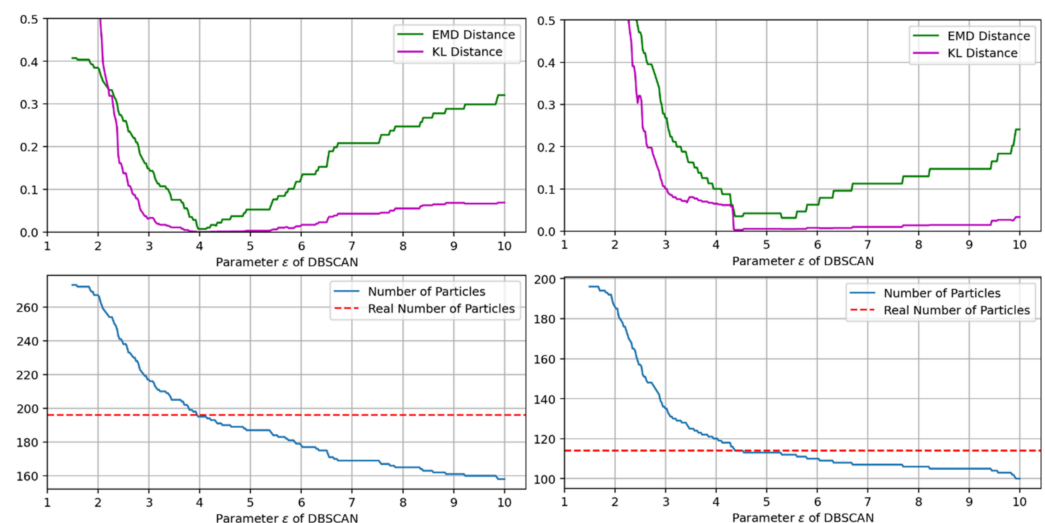


Figure 14. The dependence of the histogram difference on the radius of a neighborhood $\varepsilon = 4.33$ of the DBSCAN algorithm for images C0_0023 (left) and C0_0025 (right).

Tables 1 and 2 present clustering quality metrics for various values ε of the neighborhood radius of the DBSCAN algorithm. For an ideal arrangement of lines corresponding to nanoparticle layers, which are strictly parallel to each other and have the same length, the theoretical value of the distance function (1) is $\varepsilon (1.168 \cdot 10^{-3})$. In practice, there are cases where the lines are closer to each other. This value of $2.210 \cdot 10^{-2}$ is sufficient for the DBSCAN algorithm to place each line in a separate cluster. The tables present metric

values for a given parameter value, allowing for a comparative assessment of the quality improvement with optimal parameters for each image.

Table 1. Clustering quality metrics for C0_0023.

ε	KL Distance	EMD Metric	Particles Number	Comments
–	0	0	194	Ground Truth
1.5	2.485	0.407	273	Each line forms a separate cluster
4.125	$2.413 \cdot 10^{-4}$	$7.216 \cdot 10^{-3}$	195	Optimal value of ε for C0_0023
4.368	$1.168 \cdot 10^{-3}$	$2.210 \cdot 10^{-2}$	191	Optimal value of ε for C0_0025
4.33	$5.209 \cdot 10^{-4}$	$1.466 \cdot 10^{-2}$	192	Selected value for both images

Table 2. Clustering quality metrics for C0_0025.

ε	KL Distance	EMD Metric	Particles Number	Comments
–	0	0	114	Ground Truth
1.5	4.427	0.719	196	Each line forms a separate cluster
4.368	$2.817 \cdot 10^{-3}$	$3.509 \cdot 10^{-2}$	114	Optimal value of ε for C0_0025
4.125	$6.164 \cdot 10^{-2}$	$8.742 \cdot 10^{-2}$	118	Optimal value of ε for C0_0023
4.33	$4.034 \cdot 10^{-3}$	$4.851 \cdot 10^{-2}$	115	Selected value for both images

Based on the data presented in Tables 1 and 2, we can conclude that the optimal values (4.125 and 4.368) selected individually for each image are close and provide an acceptable result when applied to the second image. To select a single parameter, we combined the lines found in the first and second images into a single set of objects, which were then subjected to a clustering procedure. The plot of the histogram differences for the combined set of lines versus ε is shown in Figure 15.

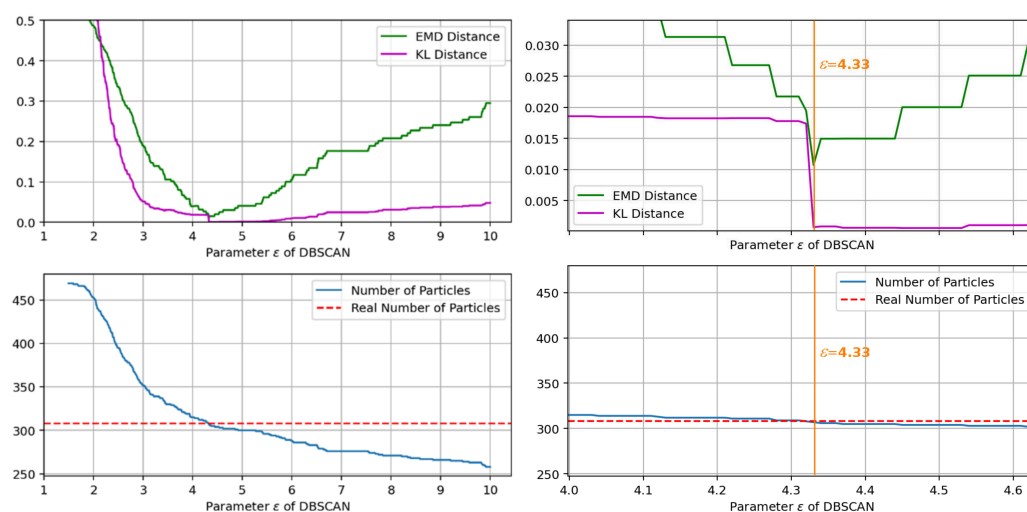


Figure 15. Dependence of histogram difference metrics on the neighborhood radius ε for combined dataset (right: zoomed view).

The value of the neighborhood radius ε , based on the analysis of the graphs presented in Figure 15, was chosen to be equal to 4.33 as the value with minimum metric magnitudes. For the combined lines of the two images, with this value of the parameter ε KL distance was $7.147 \cdot 10^{-4}$, and the EMD metric was $1.071 \cdot 10^{-2}$.

From Tables 1 and 2, it is clear that with this choice, the values of the metrics do not differ too much from the optimal ones for each image separately.

7.3. Qualitative Analysis of the Results of Combining Lines into Particles

The discrepancy between the metrics reflects the differences in the results of automatic clustering and clustering performed by an expert.

Each of the images exhibits three differences, as shown in Figure 16. For the errors in image C0_0023, the algorithm demonstrated a tendency toward excessive aggregation—individual lines were grouped or added to clusters—while for image C0_0025, on the contrary, the errors were the result of large particles being broken down into their component parts.

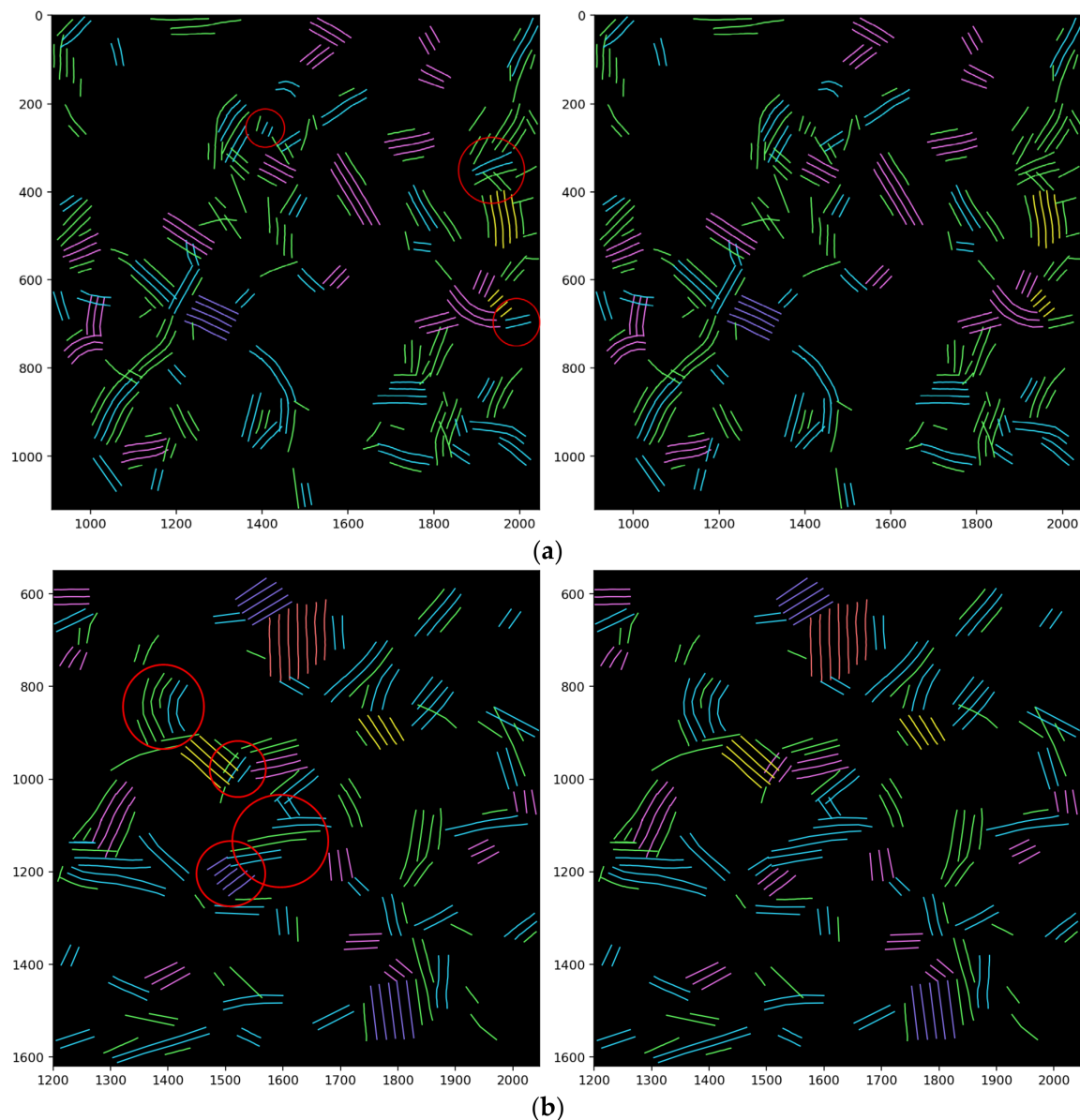


Figure 16. Results of the qualitative analysis of automatic clustering of image fragments (**left**) compared to expert labeling (**right**): (a) C0_0023 and (b) C0_0025. The color scheme of the labeling is similar to Figure 12. Differences are circled in red.

It is worth noting that, due to the labor-intensive nature of the labeling procedure, no inter-expert agreement analysis was conducted to quantify any discrepancies in judgment. We believe that the presence of only a few isolated instances of errors made by the clustering algorithm on the proposed distance function can be interpreted as a very high result.

8. Ablation Study

In Section 6, we investigated the behavior of the distance function under variations in individual line arrangement characteristics: interline distance, length ratio, mutual orientation, parallel shift, and uniform scaling. This section examines the influence of individual components of the distance function on clustering quality using real-world data.

In previous experiments, we employed two quality metrics to evaluate the correspondence between the predicted stacking number distribution and the ground-truth distribution provided by experts. Although evaluating the stacking number distribution is one of the ultimate objectives of HTERM image analysis, a more precise instrument is required for an ablation study. Consequently, in addition to the KL distance and EMD metric, we utilized the Normalized Mutual Information (NMI) [46] between the clustering generated by DBSCAN with different distance function components and the ground truth expert clustering. To align the scale of this additional quality measure with the previously used metrics, we employed the value $1 - \text{NMI}$.

Five variants of the distance function were investigated:

- Direct representation of a polyline as a set of its vertex coordinates proves inapplicable, as the vertex numbering order may vary. Therefore, as a baseline variant, we applied a component $d_{gap}(l_i, l_j)$ that monotonically depends on the minimum mean Euclidean distance between the polylines under both forward and reverse vertex numbering directions.
- Three other variants were formed by sequentially removing components $R_{len}(l_i, l_j)$, $R_{dist}(l_i, l_j)$, and $R_{diag}(l_i, l_j)$ one at a time.
- The complete distance function (1).

The experimental results are presented in Tables 3 and 4. For each function variant, the optimal neighborhood radius was selected based on the NMI metric.

Table 3. Results of the ablation study on distance function components with optimized DBSCAN neighborhood radius for image C0_0023. Values corresponding to the minimum error and the true particle number are highlighted in bold.

Distance Function Type	KL Distance	EMD Metric	1 – NMI Score	Particles Number	ε
$D(l_i, l_j) = d_{gap}(l_i, l_j) \cdot \cancel{R_{len}(l_i, l_j)} \cdot \cancel{R_{dist}(l_i, l_j)} \cdot \cancel{R_{diag}(l_i, l_j)}$	$9.92 \cdot 10^{-4}$	$1.664 \cdot 10^{-2}$	$8.686 \cdot 10^{-3}$	193	2.19
$D(l_i, l_j) = d_{gap}(l_i, l_j) \cdot \cancel{R_{len}(l_i, l_j)} \cdot R_{dist}(l_i, l_j) \cdot R_{diag}(l_i, l_j)$	$2.594 \cdot 10^{-3}$	$2.7 \cdot 10^{-2}$	$4.802 \cdot 10^{-3}$	193	3.55
$D(l_i, l_j) = d_{gap}(l_i, l_j) \cdot R_{len}(l_i, l_j) \cdot \cancel{R_{dist}(l_i, l_j)} \cdot R_{diag}(l_i, l_j)$	$5.312 \cdot 10^{-3}$	$6.846 \cdot 10^{-2}$	$6.002 \cdot 10^{-3}$	185	3.66
$D(l_i, l_j) = d_{gap}(l_i, l_j) \cdot R_{len}(l_i, l_j) \cdot R_{dist}(l_i, l_j) \cdot \cancel{R_{diag}(l_i, l_j)}$	$1.309 \cdot 10^{-4}$	$7.216 \cdot 10^{-3}$	$2.462 \cdot 10^{-3}$	194	3.44
$D(l_i, l_j) = d_{gap}(l_i, l_j) \cdot R_{len}(l_i, l_j) \cdot R_{dist}(l_i, l_j) \cdot R_{diag}(l_i, l_j)$	$4.556 \cdot 10^{-4}$	$1.031 \cdot 10^{-2}$	$2.154 \cdot 10^{-3}$	194	4.13

Table 4. Results of the ablation study on distance function components with optimized DBSCAN neighborhood radius for image C0_0025. Values corresponding to the minimum error and the true particle number are highlighted in bold.

Distance Function Type	KL Distance	EMD Metric	1 – NMI Score	Particles Number	ε
$D(l_i, l_j) = d_{gap}(l_i, l_j) \cdot \cancel{R_{len}(l_i, l_j)} \cdot \cancel{R_{dist}(l_i, l_j)} \cdot \cancel{R_{diag}(l_i, l_j)}$	$8.202 \cdot 10^{-3}$	$5.263 \cdot 10^{-2}$	$1.09 \cdot 10^{-2}$	116	2.23
$D(l_i, l_j) = d_{gap}(l_i, l_j) \cdot \cancel{R_{len}(l_i, l_j)} \cdot R_{dist}(l_i, l_j) \cdot R_{diag}(l_i, l_j)$	$4.244 \cdot 10^{-3}$	$4.851 \cdot 10^{-2}$	$7.063 \cdot 10^{-3}$	115	3.96
$D(l_i, l_j) = d_{gap}(l_i, l_j) \cdot R_{len}(l_i, l_j) \cdot \cancel{R_{dist}(l_i, l_j)} \cdot R_{diag}(l_i, l_j)$	$2.127 \cdot 10^{-3}$	$2.438 \cdot 10^{-2}$	$7.073 \cdot 10^{-3}$	113	3.36
$D(l_i, l_j) = d_{gap}(l_i, l_j) \cdot R_{len}(l_i, l_j) \cdot R_{dist}(l_i, l_j) \cdot \cancel{R_{diag}(l_i, l_j)}$	$5.483 \cdot 10^{-3}$	$4.207 \cdot 10^{-2}$	$5.433 \cdot 10^{-3}$	113	3.76
$D(l_i, l_j) = d_{gap}(l_i, l_j) \cdot R_{len}(l_i, l_j) \cdot R_{dist}(l_i, l_j) \cdot R_{diag}(l_i, l_j)$	$2.817 \cdot 10^{-3}$	$3.509 \cdot 10^{-2}$	$4.178 \cdot 10^{-3}$	114	4.37

The ablation study confirms the necessity of all proposed components in the distance function, as the complete variant consistently achieved the lowest $1 - \text{NMI}$ errors

(2.154×10^{-3} for C0_0023 and 4.178×10^{-3} for C0_0025) compared to the baseline and partial variants.

While removing individual components such as diagonal alignment or length ratio resulted in intermediate performance, the baseline variant relying solely on mean Euclidean distance demonstrated the poorest clustering fidelity, highlighting that simple spatial proximity is insufficient for accurate line arrangement analysis. Furthermore, the discrepancy between low KL divergence and higher NMI errors in partial variants underscores that while distributional metrics may appear favorable, they can mask structural clustering inaccuracies that are effectively resolved only by the full combination of geometric descriptors.

9. Discussion and Conclusions

Automated analysis of HRTEM micrographs of sulfide catalysts reduces the labor intensity of data processing and decreases the subjectivity of interpretation. Solving this problem can be divided into two main stages: identifying lines corresponding to layers and combining the identified lines into groups corresponding to sulfide nanoparticles to calculate their morphological parameters that influence catalyst activity. It should be noted that this problem is virtually unknown in the literature, and the application of currently popular deep machine learning methods is hampered by the complexity of labeling.

The paper considers the second stage of this problem within the framework of a feature-free approach, which allows working with pairwise relationships between objects (in the context of this work, between polylines) without moving to a fixed feature space.

We have proposed a combined distance function between polylines that takes into account the physicochemical features of the structure of nanoparticles:

- Geometric distance between layers with exponential normalization to the reference value;
- The ratio of line lengths to prevent layers of very different lengths from merging;
- Homogeneity of pairwise distances for parallelism control;
- The ratio of “diagonals” to identify unacceptable layer shifts.

This distance function formalizes for the first time the model representations of the texture of sulfide catalysts and at the same time satisfies the properties of a semimetric (non-negativity, identity, symmetry), which determines the applicability of density and graph clustering algorithms.

We systematically investigated the sensitivity of the distance function to center-to-center distance, aspect ratio, angular divergence, parallel shift, and scaling using model experiments. The results confirmed the monotonicity, discriminatory nature, and relative scale invariance of the function.

Experiments on real data with manual labeling by experts showed high agreement between automatic clustering and expert annotations:

- With the optimal parameter value $\varepsilon = 4.33$, the Kullback–Leibler divergence was $\sim 7 \times 10^{-4}$, and the Earth Mover’s Distance metric was $\sim 1 \times 10^{-2}$ for the combined data set.
- The number of discrepancies with expert markings did not exceed three cases per image, which indicates the practical applicability of the method.

Future publications will present a method for automatic line detection in HRTEM images and the integration of the proposed method with line detection to create an end-to-end analysis pipeline.

Overall, the developed approach demonstrates high accuracy and physical validity, making it a promising tool for materials science studies of sulfide catalysts and rational design of highly active catalytic systems.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/app16147075/s1>, CVAT project backup with HTERM images and manual expert annotations.

Author Contributions: Conceptualization, O.S., A.K., M.K and A.M.; methodology, O.S. and A.K.; software, O.K. and A.K.; validation, M.K. and M.B.; investigation, M.K.; data curation, O.K. and M.B.; writing—original draft preparation, A.K., O.S. and M.K.; writing—review and editing, M.K. and A.M.; All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by the Ministry of Science and Higher Education of the Russian Federation within the framework of the state task FEWG-2024-0001.

Data Availability Statement: The data sets used in the current study are available from the authors on reasonable request.

Acknowledgments: The sulfide catalysts and their microphotographs used in this study were obtained within the framework of the state task of Topchiev Institute of Petrochemical Synthesis RAS.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

HRTEM	High-Resolution Transmission Electron Microscopy
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
KL	Kullback–Leibler distance
EMD	Earth mover’s distance
IBN	Image-based nanocrystallography
NMI	Normalized Mutual Information

References

1. Wang, C.; Li, T.; Xu, W.; Wang, S.; Wang, K. Recent Advances in Co-Processing Biomass Feedstock with Petroleum Feedstock: A Review. *Front. Energy* **2024**, *18*, 735–759. [\[CrossRef\]](#)
2. Achour, A.; Anttila, A.; Creaser, D.; Olsson, L. Optimizing Biofuel Production: Direct Integration and Co-Hydroprocessing of Hydrolysis Lignin into Oil Refineries over an Unsupported NiMoP Catalyst. *Energy Convers. Manag.* **2025**, *327*, 119606. [\[CrossRef\]](#)
3. Alvarez-Majmutov, A.; Badoga, S.; Chen, J.; Monnier, J.; Zhang, Y. Co-Processing of Deoxygenated Pyrolysis Bio-Oil with Vacuum Gas Oil through Hydrocracking. *Energy Fuels* **2021**, *35*, 9983–9993. [\[CrossRef\]](#)
4. Eijssbouts, S.; Battiston, A.A.; van Leerdam, G.C. Life Cycle of Hydroprocessing Catalysts and Total Catalyst Management. *Catal. Today* **2008**, *130*, 361–373. [\[CrossRef\]](#)
5. Bukhtiyarova, M.V.; Vlasova, E.N.; Bukhtiyarova, G.A. Hydroprocessing of Biomass Feedstock over Sulfided CoMo-, NiMo-, and NiW-Supported Catalysts for Bio-Jet Fuel Component Production: A Review. *Sustain. Energy Fuels* **2024**, *8*, 3524–3544. [\[CrossRef\]](#)
6. Liu, J.; Li, W.Y.; Feng, J.; Gao, X. Influence of Ni on the Active Phase and Hydrodenitrogenation and Hydrosulfurization Activities of MoS₂ Catalysts. *J. Fuel Chem. Technol.* **2021**, *49*, 1513–1521. [\[CrossRef\]](#)
7. Payen, E.; Hubaut, R.; Kasztelan, S.; Poulet, O.; Grimblot, J. Morphology Study of MoS₂- and WS₂-Based Hydrotreating Catalysts by High-Resolution Electron Microscopy. *J. Catal.* **1994**, *147*, 123–132. [\[CrossRef\]](#)
8. Vollath, D.; Szabó, D.V. Synthesis of Nanocrystalline MoS₂ and WS₂ in a Microwave Plasma. *Mater. Lett.* **1998**, *35*, 236–244. [\[CrossRef\]](#)
9. Pakharukova, V.P.; Yatsenko, D.A.; Gerasimov, E.Y.; Vlasova, E.; Bukhtiyarova, G.A.; Tsybulya, S.V. Total Scattering Debye Function Analysis: Effective Approach for Structural Studies of Supported MoS₂-Based Hydrotreating Catalysts. *Ind. Eng. Chem. Res.* **2020**, *59*, 10914–10922. [\[CrossRef\]](#)
10. Maksov, A.; Dyck, O.; Wang, K.; Xiao, K.; Geohegan, D.B.; Sumpter, B.G.; Vasudevan, R.K.; Jesse, S.; Kalinin, S.V.; Ziatdinov, M. Deep Learning Analysis of Defect and Phase Evolution during Electron Beam-Induced Transformations in WS₂. *npj Comput. Mater.* **2019**, *5*, 12. [\[CrossRef\]](#)
11. Leth Larsen, M.H.; Dahl, F.; Hansen, L.P.; Barton, B.; Kisielowski, C.; Helveg, S.; Winther, O.; Hansen, T.W.; Schiøtz, J. Reconstructing the Exit Wave of 2D Materials in High-Resolution Transmission Electron Microscopy Using Machine Learning. *Ultramicroscopy* **2023**, *243*, 113641. [\[CrossRef\]](#) [\[PubMed\]](#)

12. Huang, W.; Jin, Y.; Li, Z.; Yao, L.; Chen, Y.; Luo, Z.; Zhou, S.; Lin, J.; Liu, F.; Gao, Z.; et al. Auto-Resolving the Atomic Structure at van Der Waals Interfaces Using a Generative Model. *Nat. Commun.* **2025**, *16*, 2927. [CrossRef] [PubMed]
13. Zhou, R.; Liu, H.; Babichuk, I.S.; Romaniuk, Y.A.; Tiutiunnyk, A.; Zhang, J.; Pu, Y.; Zhou, Z.; Laroze, D.; Yang, J. A Lightweight Model and Multi-Agent System for Layer Identification in Two-Dimensional Materials. *Comput. Mater. Sci.* **2025**, *259*, 114106. [CrossRef]
14. Kaur, M.; Sandino, N.T.; Terry, J.P.; Ghafariasl, M.; Abate, Y. Excitonic Landscapes in Monolayer Lateral Heterostructures Revealed by Unsupervised Machine Learning. *Adv. Opt. Mater.* **2026**, *14*, e03674. [CrossRef]
15. Xu, C.; Deng, X.; Yu, P. High-Throughput Computational Study and Machine Learning Prediction of Electronic Properties in Transition Metal Dichalcogenide/Two-Dimensional Layered Halide Perovskite Heterostructures. *ACS Appl. Mater. Interfaces* **2024**, *16*, 55970–55980. [PubMed]
16. Fraundorf, P.; Qin, W.; Moeck, P.; Mandell, E. Making Sense of Nanocrystal Lattice Fringes. *J. Appl. Phys.* **2005**, *98*, 114308. [CrossRef]
17. Zhong, J.; Meng, Y.; Liu, Z.; Zeng, F. A Novel Method for the Intelligent Recognition of Lattice Fringes in Coal HRTEM Images Based on Semantic Segmentation and Fuzzy Superpixels. *ACS Omega* **2022**, *7*, 15037–15047. [CrossRef] [PubMed]
18. Gamdha, D.; Fair, R.; Krishnamurthy, A.; Gomez, E.D.; Ganapathysubramanian, B. GRATEv2: Computational Tools for Real-Time Analysis of High-Throughput High-Resolution TEM (HRTEM) Images of Conjugated Polymers. *Mater. Adv.* **2025**, *6*, 6820–6842. [CrossRef]
19. Neumann, S.; Rafaja, D. Correlative Multi-Scale Characterization of Nanoparticles Using Transmission Electron Microscopy. *Powders* **2024**, *3*, 531–549. [CrossRef]
20. Peng, Y.Z.; Song, Y.; Xie, X.Q.; Li, Y.; Qian, Z.H.; Bai, R. Characterization of Atomic-Layer MoS₂ Synthesized Using a Hot Filament Chemical Vapor Deposition Method. *Chin. Phys. B* **2016**, *25*, 058104. [CrossRef]
21. Wang, Y.X.; Chou, S.L.; Wexler, D.; Liu, H.K.; Dou, S.X. High-Performance Sodium-Ion Batteries and Sodium-Ion Pseudocapacitors Based on MoS₂/Graphene Composites. *Chem.—A Eur. J.* **2014**, *20*, 9607–9612. [CrossRef] [PubMed]
22. Oviroh, P.O.; Jen, T.C.; Ren, J.; van Duin, A. Towards the Realisation of High Permi-Selective MoS₂ Membrane for Water Desalination. *npj Clean. Water* **2023**, *6*, 14. [CrossRef]
23. Wang, D.; Yang, C.; Huang, R.; Liu, H.; Ma, H.; Qu, W.; Tian, Z. Stability of MoS₂ Nanocatalysts for the Slurry-Phase Catalytic Hydrogenation of Anthracene. *ACS Omega* **2024**, *9*, 23843–23852. [CrossRef] [PubMed]
24. Kasztelan, S.; Toulhoat, H.; Grimblot, J.; Bonnelle, J.P. A Geometrical Model of the Active Phase of Hydrotreating Catalysts. *Appl. Catal.* **1984**, *13*, 127–159. [CrossRef]
25. Ramos, M.A.; Chianelli, R.; Enriquez-Carrejo, J.L.; Gonzalez, G.A.; Berhault, G. Metallic States by Angular Dependence in 2H-MoS₂ Slabs. *Comput. Mater. Sci.* **2014**, *84*, 18–22. [CrossRef]
26. Isshiki, T.; Nishio, K.; Saijo, H.; Shiojiri, M.; Yabuuchi, Y.; Takahashi, N. High-Resolution Transmission Electron Microscopy of Hexagonal and Rhombohedral Molybdenum Disulfide Crystals. *Microsc. Res. Tech.* **1993**, *25*, 325–334. [CrossRef] [PubMed]
27. Isshiki, T.; Nishio, K.; Aoyagi, I.; Yabuuchi, Y.; Takahashi, N.; Saijo, H.; Shiojiri, M. High Resolution Transmission Electron Microscopy of Layer Structure and Stacking Faults in Tungsten Disulphide Lubricants. *Wear* **1993**, *170*, 55–61. [CrossRef]
28. Hilton, M.R.; Fleischauer, P.D. TEM Lattice Imaging of the Nanostructure of Early-Growth Sputter-Deposited MoS₂ Solid Lubricant Films. *J. Mater. Res.* **1990**, *5*, 406–421. [CrossRef]
29. Mottl, V.; Seredin, O.; Krasotkina, O. Compactness Hypothesis, Potential Functions, and Rectifying Linear Space in Machine Learning. In *Braverman Readings in Machine Learning. Key Ideas from Inception to Current State*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2018; Volume 11100, pp. 52–102.
30. Dvoenko, S. Clustering and Separating of a Set of Members in Terms of Mutual Distances and Similarities. *Trans. Mach. Learn. Data Min.* **2009**, *2*, 80–99.
31. Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, 2–4 August 1996.
32. Kibsgaard, J.; Lauritsen, J.V.; Lægsgaard, E.; Clausen, B.S.; Topsøe, H.; Besenbacher, F. Cluster-Support Interactions and Morphology of MoS₂ Nanoclusters in a Graphite-Supported Hydrotreating Model Catalyst. *J. Am. Chem. Soc.* **2006**, *128*, 13950–13958. [CrossRef] [PubMed]
33. Niu, X.; Zhou, W.; Han, Y.; Liu, Y. Controllable Synthesis of Defect-Rich CoMoS Catalysts with Different Morphologies for the Ultradeep Hydrodesulfurization of 4,6-Dimethyldibenzothiophene. *Langmuir* **2021**, *37*, 14254–14264. [CrossRef] [PubMed]
34. Zhang, X.; Chen, B.; Wang, J.; Zhou, Y.; Huang, X. Synthesis of Loaded Monolayer MoS₂ Catalysts and Effect on Coal Tar Hydrotreating Performance. *Mater. Lett.* **2024**, *361*, 136032. [CrossRef]
35. Nogueira, A.; Znaiguia, R.; Uzio, D.; Afanasiev, P.; Berhault, G. Curved Nanostructures of Unsupported and Al₂O₃-Supported MoS₂ Catalysts: Synthesis and HDS Catalytic Properties. *Appl. Catal. A Gen.* **2012**, *429–430*, 92–105. [CrossRef]
36. Inorganic Crystal Structures Dataset (ICSD). Available online: <https://icsd.products.fiz-karlsruhe.de/> (accessed on 2 March 2026).

37. Torshin, I.Y.; Rudakov, K.V. On Metric Spaces Arising during Formalization of Recognition and Classification Problems. Part 1: Properties of Compactness. *Pattern Recognit. Image Anal.* **2016**, *26*, 274–284. [[CrossRef](#)]
38. Jain, A.K. Data Clustering: 50 Years beyond K-Means. *Pattern Recognit. Lett.* **2010**, *31*, 651–666. [[CrossRef](#)]
39. Ward, J.H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244. [[CrossRef](#)]
40. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *J. R. Stat. Soc. Ser. B (Methodol.)* **1977**, *39*, 1–22. [[CrossRef](#)]
41. Shlezinger, M.I. Spontaneous discrimination of patterns. *Read. Autom.* **1965**, 38–45.
42. Campello, R.J.G.B.; Moulavi, D.; Zimek, A.; Sander, J. Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. *ACM Trans. Knowl. Discov. Data (TKDD)* **2015**, *10*, 5. [[CrossRef](#)]
43. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [[CrossRef](#)]
44. Vaserstein, L.N. Markov Processes over Denumerable Products of Spaces, Describing Large Systems of Automata. *Probl. Peredachi Informatsii* **1969**, *5*, 64–72.
45. Rubner, Y.; Tomasi, C.; Guibas, L.J. Metric for Distributions with Applications to Image Databases. In *Proceedings of the IEEE International Conference on Computer Vision, Bombay, India, 7 January 1998*; IEEE: New York, NY, USA, 1998; pp. 59–66. [[CrossRef](#)]
46. Strehl, A.; Ghosh, J. Cluster Ensembles-A Knowledge Reuse Framework for Combining Multiple Partitions. *J. Mach. Learn. Res.* **2002**, *3*, 583–617.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.