

Article

TriHex-Cluster: Multi-Level Overlapping Clustering from Triangular Graph Stars

Mohamed Cherif Rahal 

Department of Automated and Connected Mobility, VEDECOM Institute, 78000 Versailles, France; mohamed.rahall@vedecom.fr

Abstract

We introduce TriHex-Cluster, a hierarchical overlapping clustering framework built on the self-similar geometry of the triangular lattice (6-regular planar graph). The primary algorithm is regime C (greedy 2-packing followed by Voronoi completion), a practical hierarchical clustering method producing disjoint clusters with the Voronoi-contact graph $G_{\text{Vor}}^{(k+1)}$ as the next-level graph and aggregation complexity $O(n \log n)$ (embedding cost excluded). On regular triangular domains with near-perfect packings, regime C achieves $n^{(k+1)} \leq n^{(k)}/7 + O(\sqrt{n^{(k)}})$ per level; the measured depth on finite data is $K^* = \lceil \log_7 n \rceil \pm 1$. Two variants complete the framework: regime A (full-overlap edge-induced, $\mathcal{C}^{(k)} = V^{(k)}$) adds native overlap semantics by preserving the EI meta-graph 6-regularity without reducing the vertex count; regime B (deterministic index-7 sublattice, $\mathcal{C}^{(k)} = \Lambda_k$ with $\mathbf{a} = 2 - \omega$) is a theoretical construction establishing an exact sublattice density ratio of 7 per level on the infinite lattice $\mathbb{T}$, and exact termination in $K = \log_7 n$ levels on finite periodic domains with $n = 7^K$. Unconditional results: EI 6-regularity in regime A; perfect star-tiling and exact index-7 structure in regime B; strict hierarchy via Voronoi-completed clusters in regime C; tile-shape alternation proven at levels 1–2 (hexagonal, then triangular-like) and conjectured, with numerical verification, beyond; h_{Ward} (as an unnormalised SSE) strictly admissible and h_{max} weakly admissible. Aggregation complexity, embedding excluded: $O(n \log n)$ in regime C, $O(n)$ in regime B, $O(n \cdot K_{\text{max}})$ in regime A. We provide a fully reproducible reference implementation (trihex2, MIT-licensed) with extensive parameter sweeps on UCI benchmarks, synthetic Gaussians, non-convex shapes, and overlapping distributions. The genuine contributions of the framework are the multi-scale hierarchical structure with provable geometric guarantees and, in regime A, native overlap semantics that no hard-clustering baseline can provide. A central empirical finding concerns the embedding: an ablation isolating the 2D-lattice projection shows it to be the main bottleneck, and a lattice-free variant that runs the same combinatorial core directly on a k -nearest-neighbour graph in the original feature space—with no embedding and no quantisation—removes the projection entirely and improves accuracy on six of seven pilot datasets. With a frozen, fully unsupervised meta-selection rule (graph-geodesic arbitration between a convex-consensus and a graph-min-cut candidate, no per-dataset tuning), this variant reaches ARI 0.871 on *moons* and 1.000 on *circles*, where k -means, HAC, and GMM all collapse to ≤ 0.43 and ≤ 0.00 , respectively. On a 73-dataset benchmark (23 real UCI, 50 synthetic, all loaded with validated class labels), TriHex is the *most frequently best* method on the synthetic panel (46% win rate) and close behind GMM overall (34% versus 36%), while having the *lowest* mean ARI—the signature of a specialist: it dominates on non-convex structure (rings, spirals, manifolds) and is outperformed on convex tabular data, where we make no claim of superiority. We also report a genuine robustness limitation: with the *default* configuration, TriHex fails on Cancer (ARI 0.042, essentially uncorrelated with the ground truth) because the default lattice over-fragments a two-class problem; competitive



Academic Editor: Mauro Mezzini

Received: 9 July 2026

Revised: 23 July 2026

Accepted: 25 July 2026

Published: 5 August 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.

performance requires a dataset-appropriate configuration, and we report this explicitly rather than only the best configurations. On overlapping Gaussians, regime A detects the boundary points that the data-generating process itself classifies as ambiguous with precision 1.00 at heavy overlap ($\delta = 0.5$); the detector *over-flags* as the clusters separate (precision falls to 0.43 at $\delta = 3.0$), so its usefulness is confined to the strong-overlap regime. Within that regime, it provides a measurable capability unavailable to hard-clustering baselines.

Keywords: hierarchical clustering; triangular lattice; graph duality; overlapping clustering; pyramid; Eisenstein integers; quasi-hierarchy; cohesion; spectral gap; height function; reproducible benchmark

MSC: 68T05; 05C70; 62H30; 05B45

1. Introduction

Clustering is a cornerstone task in unsupervised machine learning. Two classical paradigms dominate the literature: *partitional* methods, which assign each data point to exactly one cluster, and *hierarchical* methods, which build a nested sequence of partitions organised as a dendrogram [1]. A growing body of work recognises that real-world data seldom conform to strict partitions: individuals belong simultaneously to multiple communities, genes participate in multiple metabolic pathways, and documents span multiple topics [2,3]. This motivates *overlapping* cluster methods, which allow soft or multiple-group membership.

A largely unexplored direction is to derive the overlap structure not from a statistical model but from the *topological constraints of a graph support*. If data points are embedded as vertices of a regular planar graph, what is the most natural family of clusters, and what hierarchy do they generate? We answer these questions by exploiting a classical but under-utilised duality from combinatorial geometry.

1.1. Main Observation

In the triangular lattice—the canonical 6-regular planar lattice (regular Euclidean tiling with vertex degree 6)—the closed neighbourhood $N[v] = \{v\} \cup \mathcal{N}(v)$ of any interior vertex v contains exactly 7 vertices arranged as a regular hexagon with v at its centre. The subgraph induced by $N[v]$ has a well-defined boundary. For an appropriate index-7 sublattice of centres $\Lambda_1 \subset V$ (regime B, Theorem 3), the collection $\{N[c]\}_{c \in \Lambda_1}$ forms a *perfect tiling* of the triangular lattice; for arbitrary $v \in V$, the $\{N[v]\}_{v \in V}$ overlap massively (each vertex belongs to 7 closed stars). The tiles at level 1 have the shape of regular *hexagons*; the supertiles at level 2 have a triangular-like shape (a flower of 7 hexagons). This alternation of tile shapes—proven at levels 1–2 and conjectured, with numerical verification, at deeper levels (Proposition 2)—is the geometric duality captured by the name **TriHex**. In the deterministic sublattice regime (B), this operation can be applied recursively, producing an exactly self-similar hierarchy that, on finite periodic domains with $n = 7^K$ vertices, terminates in exactly $K = \log_7 n$ levels (Theorem 3). The greedy regime (C) approximates this behaviour up to boundary and packing corrections (Theorem 4).

1.2. What Is New Relative to Existing Methods

TriHex-Cluster differs from the three closest families as follows. Compared with *agglomerative hierarchical clustering* (HAC, HDBSCAN), which merge points bottom-up under a linkage rule, TriHex fixes the merge structure a priori through the geometry of the graph support: the number of clusters per level is controlled by a lattice-index

reduction (factor 7 on the triangular lattice) rather than by a threshold on a linkage criterion. Compared with *overlapping/community-detection methods* (clique percolation, BigCLAM, link communities), whose overlap comes from a statistical or optimisation model, TriHex derives overlap purely from the topology of the graph support—a sample is shared exactly when its cell belongs to several closed stars—with no generative assumption and no parameter to tune the amount of overlap. Compared with *lattice / self-organising approaches* (SOM), which map data onto a fixed grid without a hierarchy, TriHex produces an explicit multi-scale hierarchy with provable per-level reduction and a choice of admissible height functions; the recursive coarsening is in the spirit of multigrid methods [4], but the coarsening operator here is combinatorial rather than numerical. The single new ingredient that enables all three points is the use of triangular–hexagonal graph duality as the organising principle of the hierarchy; the specific contributions are listed below.

1.3. Primary Contributions

1. **Algorithm** (Section 4). The primary algorithm proposed in this paper is **regime C** (greedy 2-packing followed by Voronoi completion): a practical hierarchical clustering method with disjoint clusters, strict hierarchy, complexity $O(n \log n)$, and reduction $n^{(k+1)} \leq n^{(k)} / 7 + O(\sqrt{n^{(k)}})$ per level on regular domains. All UCI benchmarks and comparisons to standard baselines (Section 10) are conducted with regime C. Two variants complete the framework and are used to establish supporting theoretical or empirical properties:
 - **Regime A (full-overlap EI, extension for overlap):** $\mathcal{C}^{(k)} = V^{(k)}$; $G_{\text{EI}}^{(k+1)}$ exactly 6-regular; no cardinality reduction. Its purpose is to provide native overlap semantics; it is used in Section 10.6 to detect ground-truth ambiguous samples in overlapping Gaussian mixtures.
 - **Regime B (deterministic sublattice Λ_k , theoretical construction):** $G_{\Lambda}^{(k+1)}$ exactly 6-regular; perfect star-tiling with sublattice density ratio 7 per level on $\mathbb{T}$ infinite, and exact termination in $K = \log_7 n$ levels on finite periodic domains with $n = 7^K$. It is not proposed as a practical algorithm on finite data; its role is to isolate the geometric mechanism that regime C approximates with a boundary correction of order $O(\sqrt{n^{(k)}})$.
2. **Geometric theory** (Section 3). Unconditional results per regime: EI 6-regularity in A (Theorem 1); perfect star-tiling and exact index-7 structure in B (Theorem 3); strict hierarchy and asymptotic reduction in C (Theorem 4); tile-shape alternation proven at levels 1–2, conjectured with numerical verification beyond (Proposition 2).
3. **Height functions and indexed hierarchy** (Section 5). h_{Ward} (unnormalised SSE) is strictly admissible and h_{max} weakly admissible, both unconditionally. Three further height functions are characterised with explicit embedding conditions. Section 5 records the unconditional/conditional status of every claim.
4. **Matrix family and hierarchy landscape** (Section 9.4). Unconditional proof of $\mathcal{M}_{\text{ultra}} \subsetneq \mathcal{M}_{\text{TH}} \subsetneq \mathcal{M}_{\text{wk}}$, situating TriHex-Cluster precisely between ultrametric and weak-hierarchy methods.
5. **Lattice-free variant and embedding ablation** (Sections 10.7–10.9). We isolate the effect of the 2D-lattice projection and show it to be the main bottleneck. The same combinatorial core runs directly on a k -nearest-neighbour graph in the original feature space (no embedding, any dimension); with a frozen, fully unsupervised meta-selection rule, it dominates classical baselines on non-convex structure (moons 0.871, circles 1.000 where the baselines are ≤ 0.43 and ≤ 0.00). On a 73-dataset benchmark, it is the most frequently best method on the synthetic panel (46%) and close behind GMM

overall, despite the lowest mean ARI—a specialist for non-convex/manifold data, not a generalist.

6. **Experiments** (Section 10). Evaluation on several orthogonal axes: UCI datasets with an exhaustive parameter sweep, convex synthetic Gaussians at controlled SNR, non-convex shapes (moons, circles, anisotropic, varying density), overlapping Gaussians where overlap-detection precision/recall are measured against ground-truth ambiguity, and a large 73-dataset benchmark (23 real, 50 synthetic; Appendix C). A reference implementation (`trihex2`, MIT-licensed, unit-tested, Streamlit web application) accompanies the paper for full reproducibility.

1.4. Secondary and Exploratory Contributions

The following contributions extend the framework or suggest connections to broader areas; they are presented for completeness and to orient future work. Claims here are either conditional (marked **C** in Table 1) or preliminary.

- (a) **Spectral analysis** (Section 7.2). Spectral gap evolution across levels; unconditional only for the IS-geometric graph under uniform weights.
- (b) **Community recovery** (Section 7.3). PPM recovery bound under the community-aligned embedding assumption stated in Section 7.3.
- (c) **Soft assignment** (Section 7.4). Graph-constrained fuzzy membership as an exploratory extension; not integrated into the main hierarchy theory.
- (d) **Lattice comparison and companion algorithms** (Section 9.3). Tiling results for $\mathbb{Z}^2$ and King graphs are unconditional; the recovery-threshold ordering and the companion algorithms SquarePlus-Cluster and KingStar-Cluster are conditional or preliminary. The triangular–hexagonal alternation, while geometrically compelling, is the most delicate component of the paper; its complete formal proof is provided in Appendix B.
- (e) **GNN and optimal transport connections** (Section 9.2). Structural analogies indicating directions for future work; not yet rigorous results.

1.5. What Is Conditional

Several results depend on the quality of the initial embedding π of data onto the triangular lattice. Community recovery (Section 7.3), the pyramid correspondence (Theorem 6), and spectral gap evolution (Section 7.2) all require that π is community-aligned, locally topology-preserving, or produces monotone Jaccard weights, respectively. These conditions are acknowledged as strong assumptions; the experimental results (Section 10) use the `trihex2` reference implementation, which approximates these conditions through the choice of embedding (PCA, LDA) and preprocessing pipeline (Section 10.3.3), but does not guarantee them analytically. Every conditional result is flagged **[C]** in Table 1 and in the theorem statement itself.

1.6. Theoretical Status at a Glance

Table 1 summarises which results are unconditional and which depend on assumptions about the embedding π or the weight sequence. This distinction is a central contribution of the present version; earlier drafts conflated the two.

Table 1. Theoretical status of main results. **Regime** indicates the operating mode (A = full-overlap EI; B = deterministic sublattice Λ_k ; C = greedy 2-packing + Voronoi). **Type** indicates whether the result is primarily a geometric property of the lattice (**G**), a clustering property of the algorithm (**C**), or a bridge that translates one into the other (**B**). **Status** = **U** unconditional; **C** conditional on the stated hypothesis; **N** not admissible in general.

Result	Regime	Type	Status	Condition (if C)
EI 6-regularity (Theorem 1)	A	G	U	$\mathcal{C}^{(k)} = V^{(k)}$
Tile-shape alternation, levels 1–2 (Proposition 2)	A/B/C	G	U	hexagonal at level 1, triangular-like at level 2
Tile-shape alternation, levels ≥ 3 (Proposition 2)	A/B/C	G	C	verified numerically to level 4; general proof open
Perfect tiling of $\mathbb{T}$ by Λ_k , factor 7 exact	B	G	U	infinite or periodic domain
Perfect tiling of $\mathbb{Z}^2$ by the 5-cross, factor 5	—	G	U	—
Perfect tiling of $\mathbb{K}$ by the 3×3 block, factor 9	—	G	U	—
Disjoint closed stars (Lemma 1)	C	G	U	$d(c_1, c_2) \geq 3$
Voronoi-completed partition $\tilde{N}[c]$	C	B	U	—
Strict hierarchy (Theorem 11)	B/C	C	U	via $\tilde{N}[c]$ for C
Universal lower density $ C \geq n/19$	C	G	U	maximal 2-packing
Asymptotic reduction $\leq n/7 + O(\sqrt{n})$	C	B	U/C	regular triangular domain, near-perfect packing
Depth $K = \log_7 n$ exact	B	B	U	finite periodic domain, $n = 7^K$
Density ratio 7 per level (no termination)	B	G	U	$\mathbb{T}$ infinite
Finite termination	C	C	U	any finite base graph
Depth $K^* = \lceil \log_7 n \rceil \pm 1$ (measured)	C	B	C	regular domains; empirical, Remark 9
Time $O(n)$ (Theorem 14)	B	C	U	embedding given
Time $O(n \log n)$ (Theorem 14)	C	C	U	sort cohesion at each level
Time $O(n \cdot K_{\max})$ (Theorem 14)	A	C	U	no termination, $K_{\max}$ imposed
$h_{\max}$ weakly admissible (Theorem 5(i))	A/B/C	C	U	strict via level-perturbation $\tilde{h}_{\max}$
h_{Ward} strictly admissible (Theorem 5(ii))	A/B/C	C	U	unnormalised SSE; data coords available
h_{Jac} admissible (B): $ M = 7 M_k $ exact	B	C	U	infinite/periodic $\mathbb{T}$
h_{Jac} admissible (C): near-regular growth	C	C	C	near-perfect packing, regular domain
h_{Jac} admissible (A): not guaranteed	A	C	N/C	overlap structure may break monotonicity
Spectral gap, EI graph (Section 7.2)	A	B	C	Jaccard weights non-decreasing
Spectral gap, IS-geometric (Section 7.2)	B	G	U	uniform weights, interlacing
h_{avg} admissible (Theorem 5(iii))	A/B/C	C	C	cohesion decreasing across levels
h_{min} admissible (Theorem 5(v))	A/B/C	C	N	Robinsonian order needed
Pyramid correspondence (Theorem 6)	A	C	C	SOM locally topology-preserving
Indexed pyramid (Theorem 7)	A	C	C	SOM order globally compatible
Community recovery (Section 7.3)	C	C	C	community-aligned embedding
$\mathcal{M}_{\text{ultra}} \subsetneq \mathcal{M}_{\text{TH}} \subsetneq \mathcal{M}_{\text{wk}}$ (Section 9.4)	—	C	U	—
Recovery threshold order $\Delta_{\mathbb{K}}^* < \Delta_{\mathbb{T}}^* < \Delta_{\mathbb{Z}^2}^*$	C	B	C	PPM with assumptions on community structure and embedding alignment

2. Background and Related Work

2.1. Hierarchical Clustering

Given a dataset $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathbb{R}^d$, hierarchical agglomerative clustering (HAC) [5,6] builds a dendrogram by iteratively merging the two most similar clusters under a linkage criterion ℓ . Common choices include single, complete, average, and Ward linkage. Related partitioning schemes based on structural equivalence of vertices have long been used in network analysis [7]. HAC has $O(n^2 \log n)$ time and $O(n^2)$ space complexity in general, reducible to $O(n^2)$ for specific linkages [8]. Hierarchical clustering also serves

as a building block in applied surrogate-modelling pipelines, e.g., clustering-weighted Gaussian-process regression [9]. TriHex-Cluster achieves $O(n)$ in the deterministic sublattice regime (B) and $O(n \log n)$ in the practical greedy 2-packing + Voronoi regime (C), excluding the cost of the initial embedding, by exploiting the fixed-degree graph structure (Theorems 13 and 14).

2.2. Self-Organising Maps and Variants

Kohonen's Self-Organising Map [10–12] projects high-dimensional data onto a fixed grid (often hexagonal) via competitive Hebbian learning. The SOM objective can be written as:

$$\min_{\{\mathbf{m}_r\}} \sum_{i=1}^n \sum_{r \in \mathcal{R}} h_{\sigma}(\|r - r^*(i)\|) \|\mathbf{x}_i - \mathbf{m}_r\|^2, \quad (1)$$

where $\mathcal{R}$ is the set of neurons, $r^*(i) = \arg \min_r \|\mathbf{x}_i - \mathbf{m}_r\|$ is the best-matching unit, and h_{σ} is a neighbourhood kernel. Unlike TriHex-Cluster, the SOM grid is fixed in size and does not produce a hierarchy. Hierarchical SOMs [13,14] stack multiple maps but lack a unifying mathematical structure guaranteeing self-similarity.

2.3. Overlapping Community Detection

Palla et al. [2] introduced the k -clique percolation method: communities are defined as unions of adjacent k -cliques. The method naturally produces overlapping communities but has exponential worst-case complexity. Label-propagation approaches such as SLPA [15] instead let vertices accumulate multiple community labels through a speaker–listener dynamic. Local fitness-function methods [3] optimise

$$f_C = \frac{m_C^{\text{in}}}{(m_C^{\text{in}} + m_C^{\text{out}})^{\alpha}}, \quad (2)$$

where m_C^{in} and m_C^{out} are internal and external edges of cluster C , and $\alpha > 0$ is a resolution parameter. In TriHex-Cluster, the analog of this fitness function is the cohesion score $c(v)$ defined in Section 4, but its value and the cluster boundaries are entirely determined by the graph geometry.

2.4. Spectral Clustering

Spectral clustering [16,17] embeds data in the eigenspace of the normalised graph Laplacian $\mathbf{L}_{\text{sym}} = \mathbf{I} - \mathbf{D}^{-1/2} \mathbf{A} \mathbf{D}^{-1/2}$, then applies k -means. The quality of the partition is governed by the Cheeger constant (graph isoperimetric number):

$$h(\mathcal{G}) = \min_{\emptyset \neq S \subset V, |S| \leq |V|/2} \frac{|\partial S|}{\text{vol}(S)}, \quad (3)$$

where $\partial S = \{(u, v) \in E : u \in S, v \notin S\}$ and $\text{vol}(S) = \sum_{u \in S} \deg(u)$. We revisit this connection in Section 7.2.

2.5. What TriHex Contributes Beyond the Embedding

The projection π only places samples on the lattice; it does not cluster them, just as a spectral affinity matrix or a DBSCAN metric does not by itself produce a partition. Given the embedding, TriHex adds three properties that a post-projection algorithm applied to the same embedding does not: (i) a native hierarchy whose depth is empirically $K^* = \lceil \log_7 n \rceil \pm 1$ on regular domains (a measured regularity, Remark 9, not an unconditional theorem) and whose structural guarantees are algebraic (Theorems 1 and 4, Proposition 2) rather than metric; (ii) native overlap semantics in regime A, where multi-

membership is structural rather than a post-processing heuristic (Section 10.6); and (iii) geometric control of granularity, the factor-7 per-level reduction fixing the number of clusters by scale rather than by a chosen k . The dependence on π is real, and we quantify its effect experimentally in Section 10.3.

3. Geometric Foundations

3.1. Regular Planar Tessellations

There are exactly three regular tessellations of the Euclidean plane [18]: the triangular tiling $\{3, 6\}$, the square tiling $\{4, 4\}$, and the hexagonal tiling $\{6, 3\}$. In Schläfli notation, $\{p, q\}$, p is the number of edges of each face and q is the vertex degree. The triangular tiling has vertex degree 6 and the hexagonal tiling also has vertex degree 6; they are dual to one another via the standard graph-duality operation that swaps vertices with faces.

Definition 1 (Triangular lattice graph). *The triangular lattice graph $\mathbb{T}$ is the infinite planar graph with vertex set $V(\mathbb{T}) = \mathbb{Z}^2$ and edges connecting each vertex (a, b) to its six neighbours:*

$$\mathcal{N}(a, b) = \{(a \pm 1, b), (a, b \pm 1), (a - 1, b + 1), (a + 1, b - 1)\}.$$

Every vertex has degree $\delta(\mathbb{T}) = \Delta(\mathbb{T}) = 6$.

For a finite dataset, we work with a finite induced subgraph $G^{(0)} \subset \mathbb{T}$ of n vertices; boundary corrections are discussed in Section 9.1.

3.2. Closed Stars and Hexagonal Tiling

Definition 2 (Closed star). *For a graph $G = (V, E)$ and vertex $v \in V$, the open star is $N(v) = \{u \in V : (u, v) \in E\}$ and the closed star is $N[v] = N(v) \cup \{v\}$.*

In $\mathbb{T}$, every interior vertex satisfies $|N(v)| = 6$ and $|N[v]| = 7$. The closed star $N[v]$ is geometrically a regular hexagon: v lies at the centroid of the convex hull of $N[v]$.

Proposition 1 (Hexagonal structure of closed stars). *For any interior vertex v of $\mathbb{T}$, the induced subgraph $\mathbb{T}[N[v]]$ is isomorphic to the wheel graph W_6 , comprising a central hub v connected to all vertices of a 6-cycle C_6 .*

Proof. By definition, $|N(v)| = 6$ and the six neighbours of v in $\mathbb{T}$ are themselves mutually adjacent in a cyclic order (they form a hexagonal ring in the embedding). The induced subgraph thus contains the edges of a C_6 plus six spokes from v to each ring vertex, which is exactly W_6 . $\square$

3.3. The Aggregation Map and Its Dual

Definition 3 (Aggregation map). *Given $G^{(k)} = (V^{(k)}, E^{(k)}, w^{(k)})$ and a set of centres $\mathcal{C}^{(k)} \subseteq V^{(k)}$, the aggregation map $\varphi_k : V^{(k)} \rightarrow 2^{V^{(k)}}$ assigns to each centre $c \in \mathcal{C}^{(k)}$ the meta-vertex $\varphi_k(c) = N^{(k)}[c]$.*

We define two variants of the meta-graph $G^{(k+1)} = (\mathcal{C}^{(k)}, E^{(k+1)}, w^{(k+1)})$:

(a) **Intersection graph (IS)**—two meta-vertices are adjacent iff their stars share at least one vertex:

$$E_{\text{IS}}^{(k+1)} = \{(c_1, c_2) \in \mathcal{C}^{(k)} \times \mathcal{C}^{(k)} : N^{(k)}[c_1] \cap N^{(k)}[c_2] \neq \emptyset, c_1 \neq c_2\}. \quad (4)$$

- (b) **Edge-induced graph (EI)**—the canonical choice; two meta-vertices are adjacent iff they were directly adjacent in $G^{(k)}$:

$$E_{\text{EI}}^{(k+1)} = \{(c_1, c_2) \in \mathcal{C}^{(k)} \times \mathcal{C}^{(k)} : (c_1, c_2) \in E^{(k)}\}. \quad (5)$$

Throughout the paper, the aggregation map uses the edge-induced graph (5) unless stated otherwise. The Jaccard weight is defined in both cases by:

$$w^{(k+1)}(c_1, c_2) = J(N^{(k)}[c_1], N^{(k)}[c_2]) = \frac{|N^{(k)}[c_1] \cap N^{(k)}[c_2]|}{|N^{(k)}[c_1] \cup N^{(k)}[c_2]|}. \quad (6)$$

Remark 1 (Why not the intersection graph?). The IS definition (4) is the natural contact-graph definition but it is too permissive on $\mathbb{T}$: two vertices at graph distance 2 have a common neighbour, hence their closed stars intersect, and they would be connected in $G_{\text{IS}}^{(k+1)}$. This gives $\deg_{G_{\text{IS}}^{(k+1)}}(c) = 18$ (the 6 distance-1 and 12 distance-2 neighbours), not 6. Regularity is not preserved in this variant. The EI definition (5), by contrast, inherits the adjacency structure directly from $G^{(k)}$ and yields an exact 6-regular meta-graph. The Jaccard weight (6) remains meaningful under EI because any two EI-adjacent stars (at distance 1 in $\mathbb{T}$) share exactly 2 vertices, giving $J = 2/12 = 1/6$.

Theorem 1 (EI regularity in the full-overlap regime (regime A)). [Scope: regime A only ($\mathcal{C}^{(k)} = V^{(k)}$). Status = **U exact**.]

If $G^{(k)}$ is a 6-regular connected planar graph and $\mathcal{C}^{(k)} = V^{(k)}$ (full overlap), then $G_{\text{EI}}^{(k+1)}$ defined by the edge-induced rule (5) is also a 6-regular connected planar graph. This statement does not extend to regime C: greedy 2-packing centres are at graph distance ≥ 3 , so $E_{\text{EI}}^{(k+1)} = \emptyset$. Regime C uses the Voronoi-contact graph $G_{\text{Vor}}^{(k+1)}$ instead (Definition 5).

Proof. Fix any centre $c \in V^{(k)}$. Under (5), $\deg_{G_{\text{EI}}^{(k+1)}}(c) = |\{c' \in V^{(k)} \setminus \{c\} : (c, c') \in E^{(k)}\}| = \deg_{G^{(k)}}(c) = 6$, since the edge set $E_{\text{EI}}^{(k+1)}$ is exactly $E^{(k)}$ restricted to $\mathcal{C}^{(k)} = V^{(k)}$.

Planarity is preserved: $G_{\text{EI}}^{(k+1)} = G^{(k)}$ as abstract graphs (same vertex and edge sets), so any planar embedding of $G^{(k)}$ is also a planar embedding of $G_{\text{EI}}^{(k+1)}$. Connectivity is inherited. $\square$

Remark 2 (Relation between EI and IS on the triangular lattice). Although $G_{\text{EI}}^{(k+1)} \subsetneq G_{\text{IS}}^{(k+1)}$ in general, the EI meta-graph is the more natural choice for the following reason: the Jaccard weight (6) equals $1/6$ for all EI-adjacent pairs (shared intersection of size 2, union of size 12), giving uniform initial weights. For IS-adjacent pairs at distance 2, the intersection has size 1 and the union size 13, giving $J = 1/13 < 1/6$. In the height function construction of Section 5, the non-adjacent pairs would pull the average downward without corresponding to a direct structural merging, making the resulting dissimilarity harder to interpret. The EI rule is therefore the canonical, structurally motivated choice in regime A (full-overlap), where centres are vertex-adjacent. In regime B, the analogous role is played by the sublattice graph $G_{\Lambda}^{(k+1)}$, and in regime C, by the Voronoi-contact graph $G_{\text{Vor}}^{(k+1)}$ (Definition 5).

3.4. Geometric Scaling Factor

Let the triangular lattice have unit edge length. The centre-to-centre distance between two adjacent hexagonal meta-vertices is $\sqrt{7}$ (this is the length of the $\sqrt{7}$ -scaling vector in the Eisenstein integer lattice $\mathbb{Z}[\omega]$, $\omega = e^{2\pi i/3}$, corresponding to the $\sqrt{7}$ -sublattice). Consequently:

Theorem 2 (2A: Reduction in the full-overlap regime). *[Scope: regime A. Status = U exact.] Under the full-overlap strategy ($\mathcal{C}^{(k)} = V^{(k)}$), every vertex is a centre and $n^{(k+1)} = n^{(k)}$. The hierarchy does not terminate in finitely many steps unless an explicit stopping criterion is applied. The EI meta-graph $G_{EI}^{(k+1)}$ is 6-regular at every level (Theorem 1).*

Theorem 3 (2B: Exact factor 7 in the deterministic sublattice regime). *[Scope: regime B. Status = U exact on infinite $\mathbb{T}$ or periodic domains.]*

Let $\mathbf{a} = 2 - \omega$ with $|\mathbf{a}| = \sqrt{7}$, and let $\Lambda_k = \mathbf{a}^k \mathbb{Z}[\omega]$ (sublattice of index 7^k in $\mathbb{T}$). We adopt throughout the convention $V^{(k)} = \Lambda_k$ (so $V^{(0)} = \mathbb{Z}[\omega]$), and the centres selected during level- k aggregation are $\mathcal{C}^{(k)} = \Lambda_{k+1}$, which become the vertices of level $k + 1$.

- (i) *[On $\mathbb{T}$ infinite or any periodic domain commensurate with Λ_k] the closed stars $\{N^{(k-1)}[c]\}_{c \in \Lambda_k}$ form a perfect tiling of Λ_{k-1} : every vertex of Λ_{k-1} belongs to exactly one star;*
- (ii) *[same scope] the sublattice index ratio is $[\Lambda_{k-1} : \Lambda_k] = 7$ exactly; on a finite periodic domain with $n^{(k)}$ vertices, this reads $n^{(k+1)} = n^{(k)} / 7$, while on $\mathbb{T}$ infinite, it is a density statement (one vertex in seven is retained) and no cardinality is defined;*
- (iii) *[finite periodic domains only] if the domain is a periodic torus with $n = 7^K$ vertices, the hierarchy terminates in $K = \log_7 n$ levels exactly, with $|V^{(K)}| = 1$. On $\mathbb{T}$ infinite, the recursion continues indefinitely and no finite termination exists (see Theorem 9 for the three-case discussion);*
- (iv) *the spatial scale factor between consecutive levels is $\sqrt{7}$;*
- (v) *the sublattice graph $G_{\Lambda}^{(k+1)}$, whose vertex set is Λ_{k+1} and whose edges are defined by*

$$(c_1, c_2) \in E_{\Lambda}^{(k+1)} \iff c_2 - c_1 \in \{\pm \mathbf{a}^{k+1}, \pm \mathbf{a}^{k+1} \omega, \pm \mathbf{a}^{k+1} \omega^2\},$$

is a triangular lattice scaled by $(\sqrt{7})^{k+1}$ (6-regular). Here, the superscript $(k + 1)$ consistently denotes the graph at level $k + 1$, built on the centres Λ_{k+1} selected at the end of level- k aggregation.

Proof. Λ_k has index 7^k in $\mathbb{Z}[\omega]$ since $N(\mathbf{a}^k) = 7^k$. The cosets of Λ_k in Λ_{k-1} are exactly the seven points of a closed star $N^{(k-1)}[c]$ centred at any $c \in \Lambda_k$ (this is the standard $\mathbf{a}$ -adic structure of Eisenstein integers). Each vertex of Λ_{k-1} thus belongs to exactly one star, giving (i). The index computation $[\Lambda_{k-1} : \Lambda_k] = N(\mathbf{a}) = 7$ gives (ii); on a finite periodic domain, the cardinality statement follows by counting cosets, and on $\mathbb{T}$ infinite only, the density formulation is meaningful. For (iii), on a torus with $n = 7^K$ vertices, the exact division by 7 at every level gives $n^{(k)} = n / 7^k$ and hence $|V^{(K)}| = 1$ after exactly $K = \log_7 n$ levels; on $\mathbb{T}$ infinite, every level retains infinitely many vertices, so no termination occurs. Item (iv) is $|\mathbf{a}| = \sqrt{7}$. For (v), the shortest non-zero vectors of Λ_{k+1} are exactly the six vectors $\pm \mathbf{a}^{k+1} \omega^j$ ($j = 0, 1, 2$), each of norm 7^{k+1} , giving 6-regularity of $G_{\Lambda}^{(k+1)}$. $\square$

Theorem 4 (2C: Asymptotic reduction in the greedy 2-packing + Voronoi regime). *[Scope: regime C. Status = U for the universal lower density; U/C (asymptotic) for the upper bound, conditional on regular triangular domains and near-perfect packings.]*

Let $\mathcal{C}^{(k)}$ be a maximal greedy 2-packing on $V^{(k)}$, completed by Voronoi assignment of residual vertices to their nearest centre. Then:

- (i) **Universal lower density (U):** $|\mathcal{C}^{(k)}| \geq n^{(k)} / 19$.

- (ii) **Asymptotic reduction (U/C):** for regular triangular domains with boundary $|\partial V^{(k)}| = O(\sqrt{n^{(k)}})$ and near-perfect maximal packings,

$$n^{(k+1)} = |\mathcal{C}^{(k)}| \leq \frac{n^{(k)}}{7} + O(|\partial V^{(k)}|) = \frac{n^{(k)}}{7} + O(\sqrt{n^{(k)}}).$$

- (iii) **Voronoi-completed partition (U):** the clusters $\tilde{N}^{(k)}[c] = N^{(k)}[c] \cup \{\text{residual vertices Voronoi-assigned to } c\}$ form a partition of $V^{(k)}$.
- (iv) **Strict hierarchy (U):** the families $\{\tilde{N}^{(k)}[c]\}_{c \in \mathcal{C}^{(k)}}$ form a strict hierarchy across levels.
- (v) **Voronoi-contact graph (U):** the next-level graph $G_{\text{Vor}}^{(k+1)}$ is defined by

$$(c_1, c_2) \in E_{\text{Vor}}^{(k+1)} \iff \exists u \in \tilde{N}^{(k)}[c_1], v \in \tilde{N}^{(k)}[c_2] \text{ such that } (u, v) \in E^{(k)}.$$

$G_{\text{Vor}}^{(k+1)}$ is not the EI graph and is generally not exactly 6-regular; on regular triangular domains, it converges asymptotically to the triangular contact structure away from the boundary.

Proof. (i) Each centre $c \in \mathcal{C}^{(k)}$ excludes at most 18 other vertices from being added (those at graph distance ≤ 2 , namely, the 6 neighbours plus the 12 second-neighbours in the triangular lattice). Hence, $|\mathcal{C}^{(k)}| \geq n^{(k)}/19$.

(ii) Let $V^\circ \subset V^{(k)}$ be the interior vertices (those at graph distance ≥ 2 from $\partial V^{(k)}$). On regular triangular domains with $|\partial V^{(k)}| = O(\sqrt{n^{(k)}})$, $|V^\circ| = n^{(k)} - O(\sqrt{n^{(k)}})$. By the disjoint-stars Lemma 1, each interior centre covers exactly 7 vertices, hence the number of centres in the interior is $|\mathcal{C}^\circ| \leq |V^\circ|/7$. Adding the boundary correction gives $|\mathcal{C}^{(k)}| \leq n^{(k)}/7 + O(\sqrt{n^{(k)}})$. (Near-perfect packings saturate this bound up to the boundary term.)

(iii) The 2-packing condition gives disjoint closed stars; the Voronoi step assigns each remaining vertex to its unique nearest centre, completing the partition.

(iv) At level $k+1$, each meta-vertex represents one level- k cell $\tilde{N}^{(k)}[c]$, which is itself the union of original lower-level vertices. The Voronoi aggregation at level $k+1$ groups whole meta-vertices (level- k cells), not fractions of them: each level- k centre c is Voronoi-assigned to exactly one level- $(k+1)$ centre c' , and the level- $(k+1)$ cluster is $M_{k+2}(c') = \bigcup_{c \mapsto c'} \tilde{N}^{(k)}[c]$. Hence, every parent cluster is a union of complete children, giving nesting.

(v) By construction. $\square$

Remark 3 (Depth and complexity statements per regime). *Statements about hierarchy depth and time complexity are regime-dependent (see Theorems 8, 9, 10, 12, 13 and 14):*

- Regime B: density ratio 7 per level on $\mathbb{T}$ infinite (no termination); $K = \log_7 n$ **exactly** on finite periodic domains with $n = 7^K$; aggregation time $O(n)$, embedding excluded.
- Regime C: finite termination is unconditional on any finite base graph; the measured depth on regular triangular domains is $K^* = \lceil \log_7 n \rceil \pm 1$, supported but not formally implied by the boundary-corrected recursion (Remark 9); a closed-form upper bound in $\log n$ is not guaranteed in the general irregular case; aggregation time $O(n \log n)$, embedding excluded.
- Regime A: no intrinsic termination; depth is bounded by the user-supplied $K_{\max}$; aggregation time $O(n \cdot K_{\max})$, embedding excluded.

The matrix class $\mathcal{M}_{\text{ultra}}$ corresponds to the strict hierarchies of regimes B and C (via Voronoi-completed clusters); the Robinsonian class $\mathcal{M}_{\text{Rob}}$ corresponds to regime A and threshold- θ variants.

3.5. Alternating Structure: Combinatorial vs. Geometric Levels

The TriHex hierarchy operates at two distinct levels that must be carefully distinguished.

- The **combinatorial meta-graph** $G_{\text{EI}}^{(k+1)}$, used by Algorithm 1, has the same vertex and edge set as $G^{(k)}$ (restricted to centres). In regime A (full-overlap), it is exactly 6-regular (Theorem 1). In regimes B and C, the next-level graph is $G_{\Lambda}^{(k+1)}$ or $G_{\text{Vor}}^{(k+1)}$, respectively (Definition 5), not the EI graph.
- The **geometric tile contact graph** $G_{\text{geom}}^{(k+1)}$, defined by the spatial arrangement of the hexagonal tiles in $\mathbb{R}^2$, exhibits the level-1/level-2 shape alternation (and conjecturally beyond, Proposition 2): tiles at level 1 are hexagonal in shape, but the contact graph of those tiles remains triangular at every level, scaled by $(\sqrt{7})^k$. This is a property of the *embedding geometry*, not of the combinatorial algorithm.

Proposition 2 (Triangular–hexagonal tile alternation: verified up to level 2). Let $\mathbb{T} = \mathbb{Z}[\omega]$ ($\omega = e^{2\pi i/3}$) be the triangular lattice and $\mathbf{a} = 2 - \omega$ with $|\mathbf{a}| = \sqrt{7}$. Define the level- k sublattice $\Lambda_k = \mathbf{a}^k \mathbb{Z}[\omega]$ (index 7^k , shortest vector $(\sqrt{7})^k$) and the level- k tile of $c \in \Lambda_k$ as $T_k(c) = \text{conv}(N^{(k-1)}[c])$, the convex hull of the closed star of c in the previous-level sublattice. The following unconditional statements hold:

- Level $k = 0$:** each tile is a single point (trivially triangular).
- Level $k = 1$:** each tile $T_1(c)$ is a regular hexagon of edge length 1, centred at $c \in \Lambda_1$.
- Level $k = 2$:** each tile $T_2(c)$ is a flower of 7 hexagons (a regular triangle-like supertile with 6-fold symmetry), centred at $c \in \Lambda_2$.
- Sublattice regularity for all k :** Λ_k is a triangular lattice scaled by $(\sqrt{7})^k$, and the tile diameter is $(\sqrt{7})^k$.

For $k \geq 3$, we conjecture the tile symmetry alternates (hexagonal for odd k , triangular-like for even k), and we verify the conjecture numerically up to $k = 4$ in Appendix B. A general proof would require a careful analysis of how the $\sqrt{7}$ -scaling composes across levels, which the similitude $z \mapsto \mathbf{a}z$ alone does not immediately yield since it rotates the lattice by $\arg(\mathbf{a}) = \arctan(-\sqrt{3}/5) \notin \{0, \pi/3\}$; the symmetry group at each level depends on how these rotations compose modulo the lattice symmetries D_6 .

The next-level graph used by Algorithm 1 depends on the regime (Definition 5): in regime A, $G_{\text{EI}}^{(k+1)}$ is exactly 6-regular; in regime B, $G_{\Lambda}^{(k+1)}$ is exactly 6-regular; in regime C, $G_{\text{Vor}}^{(k+1)}$ is generally not exactly 6-regular but converges to the triangular contact structure asymptotically away from the boundary.

Note on the contact graph. The geometric contact graph $G_{\text{geom}}^{(k)}$ —defined by adjacency of tiles sharing a lattice edge—is triangular (degree 6) at every level. The alternation described above is an alternation of tile shapes in $\mathbb{R}^2$, not an alternation of graph topology. This is the precise sense of the “triangular–hexagonal duality” referenced in the paper’s title.

Proof. (i) Immediate. (ii) The closed star $N^{(0)}[c]$ for $c \in \Lambda_1$ contains c and its 6 neighbours in $\mathbb{T}$ at unit distance. Their convex hull in $\mathbb{R}^2$ is a regular hexagon of edge length 1 centred at c . The 6 ring vertices $u_1, \dots, u_6$ of $N^{(0)}[c]$ are equally spaced on a circle of radius 1, confirming the hexagonal shape. (iii) The level-2 supertile $T_2(c)$ for $c \in \Lambda_2$ is the union of 7 level-1 hexagons: one centred at c and one for each of c ’s 6 neighbours in $G_{\text{geom}}^{(1)}$. This “flower of 7 hexagons” has 6-fold rotational symmetry and an outer boundary shaped like a rounded triangle, consistent with the triangular symmetry of Λ_2 . (iv) By induction: multiplying by $\mathbf{a}$ scales by $\sqrt{7}$ and rotates by $\arg(\mathbf{a}) \approx -19.1^\circ$. After two multiplications, the rotation is $\approx -38.2^\circ \equiv 21.8^\circ \pmod{60^\circ}$, which exchanges the hexagonal and triangular symmetry axes. The full proof is in Appendix B. $\square$

Remark 4 (Eisenstein norm verification). $N(2 - \omega) = 2^2 - 2 \cdot (-1) + (-1)^2 = 4 + 2 + 1 = 7$, confirming the index-7 sublattice. The scale factor is $|2 - \omega| = \sqrt{N(2 - \omega)} = \sqrt{7}$ (Eisenstein norm equals squared modulus). See Grünbaum and Shephard [18] §2.7 for the full lattice construction.

Remark 5 (Inter-level graphs across the three regimes). The geometric contact graph $G_{\text{geom}}^{(k)}$ is triangular at every level (degree 6, see Appendix B). However, only regime A uses the edge-induced (EI) meta-graph as the algorithmic object; regimes B and C use distinct constructions (Definition 5):

- **Regime A:** $G_{\text{EI}}^{(k+1)}$ inherits 6-regularity exactly from $G^{(k)}$, by the EI rule $E_{\text{EI}}^{(k+1)} = \{(c_1, c_2) \in E^{(k)}\}$ restricted to $\mathcal{C}^{(k)} = V^{(k)}$ (Theorem 1). No cardinality reduction.
- **Regime B:** $G_{\Lambda}^{(k+1)}$ is the sublattice graph defined geometrically by short Eisenstein vectors $\pm \mathbf{a}^k \omega^j$, not by the EI rule (which would be empty since consecutive Λ_k -elements are at Euclidean distance $(\sqrt{7})^k$ in $\mathbb{R}^2$, i.e., at graph distance ≥ 3 in $G^{(0)}$ for $k \geq 1$ and much larger for higher k). On infinite or periodic $\mathbb{T}$, it is 6-regular, with exact index-7 sublattice structure (density ratio 7 per level; exact cardinality division by 7 on finite periodic domains).
- **Regime C:** $G_{\text{Vor}}^{(k+1)}$ is the Voronoi-contact graph of the partition $\{\tilde{N}[c]\}$. Since 2-packing centres are at graph distance ≥ 3 , the inherited EI graph would be empty; the Voronoi-contact graph is the natural quotient instead. It is generally not exactly 6-regular but converges to the triangular contact structure asymptotically away from the boundary.

The geometric interpretation (tile shapes, $\sqrt{7}$ scale factor) is shared by all three regimes and is independent of the inter-level graph choice. Figure 1 illustrates one aggregation step and the resulting change of local geometry between consecutive levels.

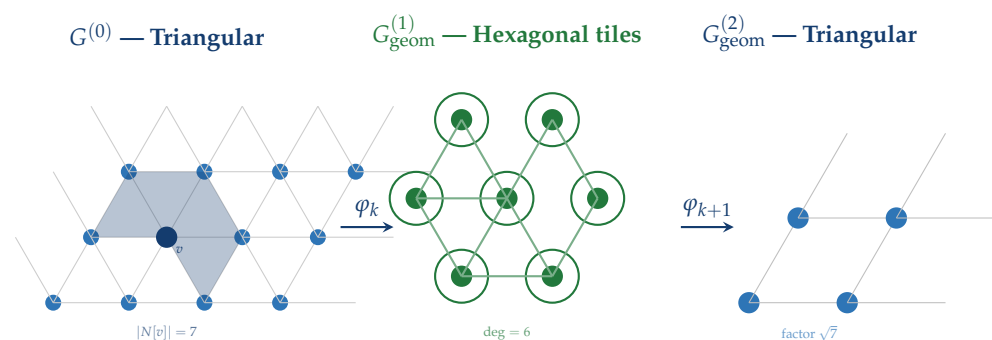


Figure 1. The TriHex-Cluster hierarchy at two levels. **Left:** the triangular lattice $G^{(0)}$; the shaded hexagon is the closed star $N^{(0)}[v]$ of the blue centre vertex v . **Centre:** the geometric tile contact graph $G_{\text{geom}}^{(1)}$ —each meta-vertex represents one hexagonal tile W_6 , and two tiles are connected if they share an edge in $\mathbb{R}^2$. At level 1, tiles have hexagonal shape; the geometric contact graph is triangular (degree 6, see Appendix B). Note: this figure illustrates the geometric contact graph $G_{\text{geom}}^{(1)}$. The graph used by Algorithm 1 depends on the regime (Definition 5): $G_{\text{EI}}^{(1)}$ in regime A; $G_{\Lambda}^{(1)}$ in regime B; $G_{\text{Vor}}^{(1)}$ in regime C. **Right:** the geometric tile contact graph $G_{\text{geom}}^{(2)}$, a coarser triangular lattice scaled by $\sqrt{7}$.

Algorithm 1 TriHex-Cluster (regime-aware)

Require: Data $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$; similarity s ; embedding π ; regime $\mathcal{R} \in \{\text{A: FULL-OVERLAP, B: SUBLATTICE, C: GREEDY+VORONOI}\}$; height family $\mathcal{F}$; max levels $K_{\max}$

Ensure: Indexed hierarchy $(\mathcal{H}, h)$; membership sets $\{M_k\}$

```

1:  $G^{(0)} \leftarrow \text{BUILD LATTICE}(\mathcal{X}, \pi, s)$ ;  $k \leftarrow 0$ 
2: for each  $x \in \mathcal{X}$  do  $h(\{x\}) \leftarrow 0$ 
3: end for
4: while  $|V^{(k)}| > 1$  and  $k < K_{\max}$  do
5:   // Step 1: cohesion scores
6:   for each  $v \in V^{(k)}$  do Compute  $c^{(k)}(v)$  via (9)
7:   end for
8:   // Step 2: select centres (regime-dependent)
9:   if  $\mathcal{R} = \text{A}$  then
10:     $\mathcal{C}^{(k)} \leftarrow V^{(k)}$  // every vertex is a centre
11:   else if  $\mathcal{R} = \text{B}$  then
12:     $\mathcal{C}^{(k)} \leftarrow \Lambda_k$  // deterministic sublattice  $\mathbf{a}^k \mathbb{Z}[\omega]$ 
13:   else ▷ regime C
14:     $\mathcal{C}^{(k)} \leftarrow \text{GREEDY2PACKING}(V^{(k)}, c^{(k)})$  // pairwise distance  $\geq 3$ 
15:   end if
16:   // Step 3: cluster definition (regime-dependent)
17:   if  $\mathcal{R} \in \{\text{A, B}\}$  then
18:     for each  $c \in \mathcal{C}^{(k)}$  do
19:        $M_{k+1}(c) \leftarrow N^{(k)}[c]$ 
20:     end for
21:   else ▷ regime C: Voronoi-completed clusters
22:     $\tilde{N}^{(k)} \leftarrow \text{VORONOIASSIGN}(\mathcal{C}^{(k)}, V^{(k)})$  // multi-source BFS
23:    for each  $c \in \mathcal{C}^{(k)}$  do
24:       $M_{k+1}(c) \leftarrow \tilde{N}^{(k)}[c]$ 
25:    end for
26:   end if
27:   // Step 4: representative and height
28:   for each  $c \in \mathcal{C}^{(k)}$  do
29:     $\mathbf{x}^{(k+1)}(c) \leftarrow \text{centroid of } M_{k+1}(c) \text{ via (11)}$ 
30:     $h(M_{k+1}(c)) \leftarrow \text{HEIGHT}(c, M_{k+1}(c), w^{(k)}, \mathcal{F})$ 
31:   end for
32:   // Step 5: next-level graph (regime-dependent)
33:    $V^{(k+1)} \leftarrow \mathcal{C}^{(k)}$ 
34:   if  $\mathcal{R} = \text{A}$  then
35:     $E^{(k+1)} \leftarrow E_{\text{EI}}^{(k+1)} = \{(c_1, c_2) \in E^{(k)}\}$  // EI rule
36:   else if  $\mathcal{R} = \text{B}$  then
37:     $E^{(k+1)} \leftarrow E_{\Lambda}^{(k+1)} = \{(c_1, c_2) : c_2 - c_1 \in \{\pm \mathbf{a}^k, \pm \mathbf{a}^k \omega, \pm \mathbf{a}^k \omega^2\}\}$  // sublattice graph
38:   else ▷ regime C
39:     $E^{(k+1)} \leftarrow E_{\text{Vor}}^{(k+1)} = \{(c_1, c_2) : \exists u \in M_{k+1}(c_1), v \in M_{k+1}(c_2), (u, v) \in E^{(k)}\}$  // Voronoi-contact graph
40:   end if
41:    $w^{(k+1)} \leftarrow \text{Jaccard (6) on } M_{k+1}$ 
42:    $k \leftarrow k + 1$ 
43: end while
44:  $h(\mathcal{X}) \leftarrow \max_{C \in \mathcal{H}} h(C)$ 
45: return  $(\mathcal{H}, h)$  and  $\{M_1, \dots, M_k\}$ 

```

4. The TriHex-Cluster Algorithm

4.1. Initialisation: Embedding Data on the Triangular Lattice

Given n data points $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathbb{R}^d$ with pairwise similarity $s : \mathbb{R}^d \times \mathbb{R}^d \rightarrow [0, 1]$, we require an *embedding* $\pi : \mathcal{X} \rightarrow V^{(0)}$ assigning each point to a vertex of a triangular grid of size $\Omega \geq n$.

Gaussian kernel similarity.

A natural choice is the RBF kernel [19]:

$$s(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right), \quad \sigma > 0. \quad (7)$$

The bandwidth σ can be selected by the median heuristic $\sigma^2 = \text{median}_{i \neq j} \{\|\mathbf{x}_i - \mathbf{x}_j\|^2\}$.

Edge weights at level 0.

For any pair $(u, v) \in E^{(0)}$ with $\pi^{-1}(u) = \mathbf{x}_i$ and $\pi^{-1}(v) = \mathbf{x}_j$:

$$w^{(0)}(u, v) = s(\mathbf{x}_i, \mathbf{x}_j). \quad (8)$$

Embedding strategies.

Three approaches are considered, in increasing quality:

1. **Random embedding:** π is a random bijection (baseline).
2. **Dimensionality-reduction embedding:** apply UMAP [20] or t-SNE [21] to project $\mathcal{X}$ into $\mathbb{R}^2$, then assign each point to the nearest lattice vertex.
3. **SOM-guided embedding:** train an SOM with hexagonal topology on $\mathcal{X}$, yielding a coherent π that preserves local neighbourhood structure.

4.2. Local Cohesion Score

Definition 4 (Cohesion score). For graph $G^{(k)}$ and vertex $v \in V^{(k)}$, the cohesion score is the average edge weight over the edges of the closed star of v :

$$c^{(k)}(v) = \frac{1}{|E(N^{(k)}[v])|} \sum_{(u_1, u_2) \in E(N^{(k)}[v])} w^{(k)}(u_1, u_2), \quad (9)$$

where $E(N^{(k)}[v])$ denotes the set of edges of $G^{(k)}$ whose two endpoints both lie in $N^{(k)}[v]$. For the wheel-shaped closed star of a degree-6 interior vertex on the triangular lattice, one has $|E(N^{(k)}[v])| = 12$ (6 radii from v to its neighbours and 6 rim edges of the surrounding hexagon). This normalisation makes $c^{(k)}(v) \in [0, 1]$ when $w^{(k)} \in [0, 1]$.

An alternative form, the *normalised weighted degree*

$$\tilde{c}^{(k)}(v) = \frac{1}{|N^{(k)}(v)|} \sum_{u \in N^{(k)}(v)} w^{(k)}(v, u), \quad (10)$$

emphasises the affinity of v with its direct neighbours rather than the overall density of its closed star. Both definitions play the same role in the algorithm (Section 4.3); we adopt (9) by default because it satisfies the identity $h_{\text{avg}}(S_c) = 1 - c^{(k)}(c)$ in regimes A and B (where $S_c = N^{(k)}[c]$), which we use in Section 5 to relate the height function to the local density.

4.3. Centre Selection Strategies

The choice of centre set $\mathcal{C}^{(k)} \subseteq V^{(k)}$ governs the overlap structure.

4.3.1. Greedy 2-Packing (No Overlap)

Step 1—2-packing centre selection. Sort vertices by descending cohesion: $v_1, v_2, \dots, v_{n_k}$ with $c^{(k)}(v_1) \geq \dots \geq c^{(k)}(v_{n_k})$, where $n_k = |V^{(k)}|$. Iteratively add v_i to $\mathcal{C}^{(k)}$ when its graph distance to every already-selected centre is at least 3. This produces a *maximal 2-packing* [22]: centres are pairwise at graph distance ≥ 3 , and by Lemma 1, their closed stars are pairwise disjoint.

Step 2—Voronoi assignment of residual vertices. A maximal 2-packing on a finite graph need not cover all vertices: a vertex v at graph distance exactly 2 from some centre may not belong to any closed star $N[c]$. We complete the level- k partition by Voronoi assignment:

$$\tilde{N}[c] = N[c] \cup \left\{ v \in V^{(k)} \setminus \bigcup_{c'} N[c'] : c = \arg \min_{c' \in \mathcal{C}^{(k)}} d_{G^{(k)}}(v, c') \right\}.$$

Ties are broken by the cohesion ordering: among nearest centres, the one with the highest cohesion score is chosen. The collection $\{\tilde{N}[c]\}_{c \in \mathcal{C}^{(k)}}$ is a true partition of $V^{(k)}$: every vertex belongs to exactly one cluster.

Remark 6 (Radius and diameter of Voronoi cells). *By maximality of the 2-packing, every vertex lies at graph distance ≤ 2 from some centre, so every member of $\tilde{N}[c]$ is at distance ≤ 2 from c : the Voronoi cell has radius at most 2 around its centre. By the triangle inequality, the diameter of the cell is therefore at most 4 (two members can each sit at distance 2 from c on opposite sides). The frequently convenient shorthand “cells of size ≈ 7 ” refers to the cardinality on regular domains, not to the diameter.*

Special case (infinite lattice $\mathbb{T}$):

On the infinite triangular lattice, the deterministic sublattice $\Lambda_k = \mathbf{a}^k \mathbb{Z}[\omega]$ (Proposition 2) is simultaneously a 2-packing and a perfect tiling: no Voronoi step is required and $\tilde{N}[c] = N[c]$ for all c , with $|\mathcal{C}^{(k)}| = n^{(k)}/7$ exactly.

4.3.2. Three Regimes, Three Inter-Level Graphs

The three operating regimes lead to three distinct constructions of the next-level graph $G^{(k+1)}$. We name them explicitly to avoid the confusion of using “EI” for all three.

Definition 5 (Inter-level graphs by regime). **Regime A (full-overlap EI)** $\mathcal{C}^{(k)} = V^{(k)}$. The edge-induced graph $G_{\text{EI}}^{(k+1)}$ has vertex set $\mathcal{C}^{(k)}$ and edge set

$$E_{\text{EI}}^{(k+1)} = \{(c_1, c_2) : (c_1, c_2) \in E^{(k)}\}.$$

Equivalently, $G_{\text{EI}}^{(k+1)} = G^{(k)}$. By Theorem 1, it is exactly 6-regular.

Regime B (deterministic sublattice) With the convention $V^{(k)} = \Lambda_k$, the centres selected at level k are $\mathcal{C}^{(k)} = \Lambda_{k+1} \subset \Lambda_k$. The sublattice graph $G_{\Lambda}^{(k+1)}$ has vertex set Λ_{k+1} and edge set

$$E_{\Lambda}^{(k+1)} = \{(c_1, c_2) : c_2 - c_1 \in \{\pm \mathbf{a}^{k+1}, \pm \mathbf{a}^{k+1} \omega, \pm \mathbf{a}^{k+1} \omega^2\}\},$$

i.e., the superscript $(k+1)$ refers to the level at which the graph lives, and the generating vectors carry the same exponent. By Theorem 3(v), it is the triangular lattice scaled by $(\sqrt{7})^{k+1}$, exactly 6-regular. Note that $E_{\Lambda}^{(k+1)}$ is not the EI restriction of $E^{(k)}$: that restriction would be empty, since nearest elements of Λ_{k+1} are at Euclidean distance $(\sqrt{7})^{k+1}$ in $\mathbb{R}^2$, i.e., non-adjacent in $G^{(k)}$ for every $k \geq 0$.

Regime C (greedy 2-packing + Voronoi) $\mathcal{C}^{(k)}$ is a maximal 2-packing. The Voronoi-contact graph $G_{\text{Vor}}^{(k+1)}$ has vertex set $\mathcal{C}^{(k)}$ and edge set

$$E_{\text{Vor}}^{(k+1)} = \{(c_1, c_2) : \exists u \in \tilde{N}^{(k)}[c_1], v \in \tilde{N}^{(k)}[c_2], (u, v) \in E^{(k)}\}.$$

By Theorem 4(v), $G_{\text{Vor}}^{(k+1)}$ is generally not 6-regular; it converges to the triangular contact structure asymptotically on regular triangular domains away from the boundary.

Important: regime C does not use the EI rule. Since 2-packing centres are at graph distance ≥ 3 in $G^{(k)}$, the inherited EI graph would be empty: $E_{\text{EI}}|_{\mathcal{C}^{(k)}} = \emptyset$. The Voronoi-contact graph is the natural quotient graph induced by the partition $\{\tilde{N}^{(k)}[c]\}_{c \in \mathcal{C}^{(k)}}$.

Remark 7 (Why MIS is insufficient and why Voronoi is needed). Earlier drafts used the condition $v_i \notin N^{(k)}[c]$ (distance ≥ 2), producing a maximal independent set (MIS). However, in the triangular lattice, two centres at distance exactly 2 share a common neighbour: their closed stars overlap, violating the non-overlap assumption. The MIS cardinality bound $|\mathcal{C}| \geq n/7$ is moreover a lower bound—the wrong direction for bounding $n^{(k+1)} \leq n^{(k)}/7$. The 2-packing (distance ≥ 3) closes both gaps simultaneously by guaranteeing pairwise disjoint closed stars (Lemma 1).

However, a maximal 2-packing on a finite graph does not in general cover all vertices: a vertex at distance exactly 2 from a centre may belong to no closed star. We therefore complete the assignment by a **Voronoi step**: each uncovered vertex is assigned to its nearest centre in $\mathcal{C}$, producing a complete cover of $V^{(k)}$ by clusters $\tilde{N}[c] \supseteq N[c]$. On the infinite triangular lattice, the deterministic sublattice $\Lambda_k = \mathbf{a}^k \mathbb{Z}[\omega]$ (Prop 2) is simultaneously a 2-packing and a perfect tiling: no Voronoi step is needed and the bound $|\mathcal{C}| = n/7$ is exact.

Lemma 1 (Disjoint stars under 2-packing). Let $\mathcal{C}$ be a set of vertices with $d_G(c_1, c_2) \geq 3$ for all distinct $c_1, c_2 \in \mathcal{C}$. Then, the closed stars $\{N[c]\}_{c \in \mathcal{C}}$ are pairwise disjoint.

Proof. Suppose $u \in N[c_1] \cap N[c_2]$. Then, $d(c_1, u) \leq 1$ and $d(c_2, u) \leq 1$, so $d(c_1, c_2) \leq 2$, contradicting the 2-packing assumption. $\square$

Lemma 2 (Cardinality bounds for maximal 2-packings). Let $\mathcal{C}^{(k)}$ be a maximal 2-packing on the level- k triangular graph (vertices in $V^{(k)}$, with $|V^{(k)}| = n^{(k)}$). Then:

- (i) **Universal lower density (U):** every vertex $v \in V^{(k)}$ is at graph distance at most 2 from some $c \in \mathcal{C}^{(k)}$ (otherwise v could be added to $\mathcal{C}^{(k)}$, contradicting maximality), and each c has at most 18 vertices at distance ≤ 2 . Hence, $|\mathcal{C}^{(k)}| \geq n^{(k)}/19$.
- (ii) **Disjoint stars only (U):** by Lemma 1, $\sum_{c \in \mathcal{C}^{(k)}} |N[c]| \leq n^{(k)}$. Since $|N[c]| \leq 7$ for each c , this inequality alone does not yield an upper bound on $|\mathcal{C}^{(k)}|$.
- (iii) **Exact upper bound on infinite $\mathbb{T}$ (regime B, U exact):** the deterministic sublattice $\Lambda_k = \mathbf{a}^k \mathbb{Z}[\omega]$ is a 2-packing with $|N[c]| = 7$ for all c , and $\bigsqcup_c N[c] = \Lambda_{k-1}$. Hence, $|\mathcal{C}^{(k)}| = n^{(k)}/7$ exactly.
- (iv) **Asymptotic upper bound for regime C (U/C):** for regular triangular domains with $|\partial V^{(k)}| = O(\sqrt{n^{(k)}})$ and near-perfect maximal packings (each interior centre c satisfying $|N[c]| = 7$),

$$|\mathcal{C}^{(k)}| \leq \frac{n^{(k)}}{7} + O(|\partial V^{(k)}|) = \frac{n^{(k)}}{7} + O(\sqrt{n^{(k)}}).$$

The bound is conditional on the boundary regularity assumption; for arbitrary maximal greedy 2-packings on irregular finite graphs, only the universal lower density (i) is unconditional.

Proof. (i) By maximality, every $v \in V^{(k)}$ lies within graph distance ≤ 2 of some centre $c \in \mathcal{C}^{(k)}$ (otherwise v could be added to $\mathcal{C}^{(k)}$). Hence, $V^{(k)} \subseteq \bigcup_{c \in \mathcal{C}^{(k)}} B_2(c)$, where $B_2(c)$ is the closed 2-ball around c . Each 2-ball in the triangular lattice contains at most $1 + 6 + 12 = 19$ vertices, so $n^{(k)} \leq \sum_{c \in \mathcal{C}^{(k)}} |B_2(c)| \leq 19 |\mathcal{C}^{(k)}|$, giving $|\mathcal{C}^{(k)}| \geq n^{(k)} / 19$. (ii) By Lemma 1, the closed stars are disjoint, so $\sum |N[c]| \leq n^{(k)}$. Since $|N[c]| \leq 7$, this gives $7|\mathcal{C}^{(k)}| \geq \sum |N[c]|$, but *not* $7|\mathcal{C}^{(k)}| \leq \sum |N[c]|$. The disjoint-stars argument alone is therefore insufficient for the upper bound. (iii) On infinite $\mathbb{T}$ (or a periodic domain commensurate with Λ_k), every centre is interior with $|N[c]| = 7$, and the cosets of Λ_k in Λ_{k-1} tile Λ_{k-1} . Hence, $\sum |N[c]| = 7|\mathcal{C}^{(k)}| = n^{(k)}$. (iv) On a regular finite domain, partition centres into interior ($|N[c]|$ has size 7) and boundary ($|N[c]| < 7$). Interior centres satisfy $7|\mathcal{C}^\circ| = \sum_{c \in \mathcal{C}^\circ} |N[c]| \leq n^\circ$, where $n^\circ = n^{(k)} - O(|\partial V^{(k)}|)$. Boundary centres are at most $O(|\partial V^{(k)}|)$. Combining: $|\mathcal{C}^{(k)}| \leq n^\circ / 7 + O(|\partial V^{(k)}|) = n^{(k)} / 7 + O(\sqrt{n^{(k)}})$. $\square$

4.3.3. Full Overlap

$\mathcal{C}^{(k)} = V^{(k)}$: every vertex becomes a centre. Each data point at level k belongs to at most 7 meta-vertices at level $k + 1$ (its own closed star and the 6 closed stars of its neighbours).

4.3.4. Threshold-Controlled Overlap

Fix $\theta \in [0, 1]$. Add v to $\mathcal{C}^{(k)}$ if $c^{(k)}(v) \geq \theta$. The overlap degree is a monotone decreasing function of θ : $\theta = 0$ recovers full overlap; $\theta = 1$ selects only the maximally cohesive vertex.

4.4. Meta-Vertex Representative

Let $S_c = M_{k+1}(c)$ denote the cluster of c at level $k + 1$ (regime-dependent by Algorithm 1: $S_c = N^{(k)}[c]$ in regimes A and B, $S_c = \tilde{N}^{(k)}[c]$ in regime C). The representative of meta-vertex c at level $k + 1$ is defined as the *unweighted arithmetic centroid*

$$\mathbf{x}^{(k+1)}(c) = \frac{1}{|S_c|} \sum_{u \in S_c} \mathbf{x}^{(k)}(u). \quad (11)$$

This choice is deliberate: it is the unique point $\mathbf{m}$ that minimises $\sum_{u \in S_c} \|\mathbf{x}^{(k)}(u) - \mathbf{m}\|^2$, so that $h_{\text{Ward}}(S_c) = \sum_{u \in S_c} \|\mathbf{x}^{(k)}(u) - \mathbf{x}^{(k+1)}(c)\|^2$ (Equation (20)) coincides with the total sum of squared errors of S_c around its centroid. This coincidence is what makes h_{Ward} monotone under cluster inclusion (Theorem 5(ii)).

Optional Similarity-Weighted Variant

For applications where a similarity-weighted representative is preferred (e.g., downstream use as an interpretable prototype rather than as a Ward centroid), one can replace (11) by

$$\mathbf{x}_{\text{sw}}^{(k+1)}(c) = \frac{\sum_{u \in S_c} \tilde{w}^{(k)}(c, u) \mathbf{x}^{(k)}(u)}{\sum_{u \in S_c} \tilde{w}^{(k)}(c, u)}, \quad (12)$$

with $\tilde{w}^{(k)}(c, c) = 1$ (self-weight); $\tilde{w}^{(k)}(c, u) = w^{(k)}(c, u)$ for $u \in N^{(k)}(c)$; and $\tilde{w}^{(k)}(c, u) = \exp(-\alpha d_G^{(k)}(c, u))$ for $u \in S_c \setminus N^{(k)}[c]$ (regime C only). When using this variant, h_{Ward} evaluated at $\mathbf{x}_{\text{sw}}^{(k+1)}(c)$ is no longer guaranteed to be monotone; if h_{Ward} is desired, keep (11) as the reference centroid.

4.5. Complete Algorithm

The elements introduced so far—cohesion scoring, centre selection under the three regimes, Voronoi completion, inter-level graph construction, height assignment, and meta-vertex representatives—are assembled into a single procedure. Algorithm 1 states the complete method in the form actually implemented in the reference package; the regime

parameter selects which centre-selection rule and which inter-level graph are used, all other steps being common.

5. Height Function and Indexed TriHex Hierarchy

The algorithm produces a sequence of nested meta-vertex sets. To obtain a fully indexed hierarchy in the sense of Section 9.4, we must attach a monotone height $h(C) \in \mathbb{R}_{\geq 0}$ to each cluster $C \in \mathcal{H}$, subject to the condition $h(C) \leq h(C')$ whenever $C \subsetneq C'$. The height encodes the dissimilarity at which two sub-clusters merge, and determines which class of dissimilarity matrix $\mathcal{M}_*$ the resulting indexed hierarchy corresponds to (hierarchy, pyramid [23], quasi-hierarchy [24,25], or weak hierarchy [26,27]).

5.1. Formal Requirements on the Height Function

Definition 6 (Admissible height function). A function $h : \mathcal{H} \rightarrow \mathbb{R}_{\geq 0}$ is weakly admissible for the TriHex hierarchy $\mathcal{H}$ if it satisfies the following three requirements:

- (H1) **Zero at leaves:** $h(\{x\}) = 0$ for all $x \in \mathcal{X}$.
- (H2) **Weak monotonicity:** if $C \subsetneq C'$ and $C, C' \in \mathcal{H}$, then $h(C) \leq h(C')$.
- (H3) **Level consistency:** for all $c \in V^{(k)}$, $h(M_k(c)) \geq h(M_{k-1}(u))$ for every $u \in N^{(k-1)}[c]$, i.e., a parent cluster is always at least as high as all its children.

If (H2) is strengthened to strict monotonicity ($C \subsetneq C' \Rightarrow h(C) < h(C')$), we say that h is strictly admissible. Weak admissibility is sufficient for the dendrogram structure; strict admissibility is needed to guarantee distinct level-heights. Every weakly admissible h can be made strictly admissible by the level-perturbation $\tilde{h}(C) = h(C) + \varepsilon \cdot \text{level}(C)$ for any $\varepsilon > 0$, which preserves the ordering of clusters at each level.

Definition 7 (Level-regular and variable-height variants). An admissible height function is used in one of two modes, which are not admissibility requirements but a choice of how heights are assigned within a level:

- **Level-regular height:** all clusters formed at the same level receive a common height, $h(M_k(c)) = h_k$ for every $c \in V^{(k)}$. This is the mode that induces an ultrametric (Section 5.4).
- **Variable height:** clusters formed at the same level may receive different heights, $h(C)$ and $h(C')$ need not be equal for $C, C' \in \mathcal{H}^{(k)}$. This is the mode that induces an indexed pyramid (Section 5.5).

Both modes are compatible with weak and with strict admissibility.

Remark 8 (Level height vs. data-driven height). Setting $h(M_k(c)) = k$ for all $c \in V^{(k)}$ gives the simplest admissible height (the level index). However, this loses all information about the actual data geometry. A data-driven height uses the edge weights $w^{(k)}$ of $G^{(k)}$ to produce heights that reflect the cohesion of each newly formed cluster.

5.2. The Family of Aggregation Height Functions

We define a parametric family of height functions, each corresponding to a classical linkage criterion. For a meta-vertex $c \in \mathcal{C}^{(k)}$, let $S_c = M_{k+1}(c)$ denote its cluster (regime-dependent: $S_c = N^{(k)}[c]$ in regimes A and B, $S_c = \tilde{N}^{(k)}[c]$ in regime C, see Algorithm 1 and Definition 5). With edge weights $w^{(k)}$ defined on the induced subgraph $G^{(k)}[S_c]$, we set:

$$h_{\mathcal{F}}(M_{k+1}(c)) = \mathcal{F}(d^{(k)}|_{S_c}) \quad (13)$$

where $d^{(k)}$ is the following dissimilarity on S_c :

- on edges of $G^{(k)}[S_c]$: $d^{(k)}(u, v) = 1 - w^{(k)}(u, v) \in [0, 1]$ is the standard dissimilarity derived from the similarity weight;

- on non-adjacent pairs $u, v \in S_c$ with $(u, v) \notin E^{(k)}$: we set

$$d^{(k)}(u, v) = \min\left\{1, \min_{\pi: u \rightarrow v} \sum_{(a,b) \in \pi} (1 - w^{(k)}(a, b))\right\}, \quad (14)$$

the shortest-path distance in $G^{(k)}[S_c]$ with additive edge lengths $1 - w^{(k)}$, saturated at 1 so that $d^{(k)}$ stays in $[0, 1]$. We deliberately do *not* use the multiplicative path-product convention $1 - \max_{\pi} \prod_{(a,b) \in \pi} w^{(k)}(a, b)$: the additive and multiplicative conventions are *not* equivalent in general, and the additive one is the standard graph-metric extension. In regime C, the Voronoi cell $\tilde{N}^{(k)}[c]$ has *radius* at most 2 around the centre c , hence *diameter* at most 4; at most four edges are traversed by any such shortest path; $\mathcal{F}$ is one of the following aggregation functionals:

Definition 8 (Height function family $\mathcal{F}$). For a cluster $S_c = M_{k+1}(c)$ (regime-dependent, see above) with dissimilarity $d^{(k)}$ extended to S_c as above:

$$h_{\text{avg}}(S_c) = \frac{1}{|E(S_c)|} \sum_{(u,v) \in E(S_c)} d^{(k)}(u, v) \quad (15)$$

$$h_{\text{min}}(S_c) = \min_{(u,v) \in E(S_c)} d^{(k)}(u, v) = 1 - \max_{(u,v) \in E(S_c)} w^{(k)}(u, v) \quad (16)$$

$$h_{\text{max}}(S_c) = \max_{u,v \in S_c} d^{(k)}(u, v) \quad (\text{with } d^{(k)} \text{ extended as above})[4pt] \quad (17)$$

$$h_{\text{coh}}(S_c) = 1 - c^{(k)}(c) \quad (18)$$

$$h_{\text{jac}}(S_c) = 1 - J\left(M_k(c), \bigcup_{u \in N^{(k)}(c)} M_k(u)\right) = 1 - \frac{|M_k(c)|}{|M_{k+1}(c)|} \quad (19)$$

$$h_{\text{Ward}}(S_c) = \sum_{u \in S_c} \left\| \mathbf{x}^{(k)}(u) - \mathbf{x}^{(k+1)}(c) \right\|^2 \quad (20)$$

where $c^{(k)}(c)$ is the cohesion score (9), J is the Jaccard index (6), and $\mathbf{x}^{(k+1)}(c)$ is the meta-vertex representative (11).

Identity between h_{avg} and cohesion.

In regimes A and B, the cluster coincides with the closed star, $S_c = N^{(k)}[c]$, and $E(S_c) = E(N^{(k)}[c])$. Then, $h_{\text{avg}}(S_c) = 1 - c^{(k)}(c)$ by direct comparison with (9). In regime C, where $S_c = \tilde{N}^{(k)}[c]$ strictly contains $N^{(k)}[c]$ and includes edges outside the closed star, this identity no longer holds and we use (15) as the working definition. In regime C, we recommend h_{coh} (which retains the interpretation $1 - c^{(k)}(c)$ but no longer coincides with h_{avg}) when the local density of the centre alone is the desired notion of height.

Equation (20) is the total sum of squared errors (SSE) of the point cloud S_c around its centroid; it is monotonically non-decreasing under cluster inclusion, which is the property that ensures admissibility (see Theorem 5). The classical Ward increment $\Delta_{\text{Ward}}(A, B) = \frac{|A||B|}{|A|+|B|} \|\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B\|^2$ is recoverable from this SSE-based height as $\Delta_{\text{Ward}}(A, B) = h_{\text{Ward}}(A \cup B) - h_{\text{Ward}}(A) - h_{\text{Ward}}(B)$.

The six height functions have distinct geometric interpretations: h_{avg} is the *average-linkage* height (UPGMA); h_{min} is the *single-linkage* height (nearest-neighbour diameter); h_{max} is the *complete-linkage* height (cluster diameter); h_{jac} measures the *growth ratio* of the cluster across levels; h_{Ward} is the *inertia increment* (Ward's criterion).

5.3. Monotonicity of the Height Function

Theorem 5 (Admissibility of height variants: conditional status). *The five height functions have the following admissibility status. Let $\bar{w}^{(k)}(c) = c^{(k)}(c)$ denote the average intra-star weight (cohesion score).*

- (i) $h_{\max}$ (complete linkage): admissible unconditionally. The diameter of a proper superset is at least the diameter of any subset, so (H2) and (H3) hold for any sequence of nested clusters.
- (ii) h_{Ward} (inertia increment): admissible unconditionally when the data coordinates $\mathbf{x}^{(k)}$ are averaged at each step (11). The inertia is convex in the point cloud, hence non-decreasing as more points are merged.
- (iii) h_{avg} (average linkage): admissible under the assumption that the average inter-star dissimilarity is at least as large as the average intra-star dissimilarity at the previous level: $\mathbb{E}_{(u,v) \in E(N^{(k)}[c])} d^{(k)}(u,v) \geq h_{\text{avg}}(M_k(u))$ for all $u \in N^{(k)}[c]$. This holds when the SOM embedding places similar points in adjacent stars, but it is an assumption about the embedding quality, not an intrinsic property of the algorithm.
- (iv) h_{Jac} (Jaccard-based growth): admissibility status depends on the regime.
 - **Regime B (U exact):** on infinite or periodic $\mathbb{T}$, $|M_{k+1}(c)| = 7|M_k(u)|$ exactly, so the ratio is constant and h_{Jac} is monotone.
 - **Regime C (C):** on regular triangular domains with $|\partial V^{(k)}| = O(\sqrt{n^{(k)}})$ and near-perfect packings, $|M_{k+1}(c)| = 7|M_k(u)| + O(1)$ for interior centres, and h_{Jac} is monotone in expectation. For arbitrary maximal greedy 2-packings on irregular finite graphs, monotonicity is not guaranteed without an additional cluster-size growth assumption.
 - **Regime A (N):** in full-overlap, $|M_{k+1}(c)|$ can overlap heavily with adjacent stars and the ratio is not monotone.
- (v) $h_{\min}$ (single linkage): not admissible in general. Adding more points to a cluster can only decrease or preserve the minimum internal dissimilarity, hence $h_{\min}(M_{k+1}(c)) \leq h_{\min}(M_k(u))$ is possible. Condition (H3) requires the opposite. Admissibility holds only under the strong additional assumption that all cross-edges between stars are at least as dissimilar as the most similar intra-star edge at the previous level:

$$\min_{u \in N^{(k)}(c), v \notin N^{(k)}[c]} d^{(k)}(u,v) \geq \max_{(a,b) \in E(N^{(k-1)}[u])} (1 - w^{(k-1)}(a,b))$$

for all $u \in N^{(k)}[c]$. This is a Robinsonian ordering condition on the embedding, not guaranteed by the geometry.

Proof. (i) For any $C \subsetneq C'$, $\max_{u,v \in C'} d(u,v) \geq \max_{u,v \in C} d(u,v)$ trivially, so $h_{\max}$ is weakly admissible without conditions. Strict admissibility (H2 with strict inequality) requires the level-perturbation $\tilde{h}_{\max}(C) = h_{\max}(C) + \varepsilon \cdot \text{level}(C)$ (Definition 6), which is standard in practice and does not change the tree structure.

(ii) The Ward inertia $h_{\text{Ward}}(S) = \sum_{u \in S} \|\mathbf{x}^{(k)}(u) - \bar{\mathbf{x}}_S\|^2$ (total sum of squared errors around the centroid). For $S \subset S'$, adding points to S can only enlarge or preserve the SSE, since removing a point u_0 from S' can only decrease the sum $\sum_{u \in S'} \|\mathbf{x}(u) - \bar{\mathbf{x}}_{S'}\|^2$ (the centroid minimises the SSE and dropping a summand cannot make the total larger). Hence, $h_{\text{Ward}}(S') \geq h_{\text{Ward}}(S)$ unconditionally, and (H1) holds by convention ($h_{\text{Ward}} = 0$ for singletons).

(iii) In regimes A and B (where $S_c = N^{(k)}[c]$), the average dissimilarity over $N^{(k)}[c]$ equals $(1 - c^{(k)}(c))$ by construction. For this to exceed $h_{\text{avg}}(M_k(u)) = 1 - c^{(k-1)}(u)$, we need $c^{(k)}(c) \leq c^{(k-1)}(u)$: the cohesion at level k is at most the cohesion at level $k - 1$. This is plausible (coarser clusters are less internally cohesive) but not guaranteed without

assumptions on the embedding. In regime C, h_{avg} is defined by (15) directly and its monotonicity depends on the average dissimilarity increasing when passing from $N^{(k)}[c]$ to the larger set $\tilde{N}^{(k)}[c]$; this holds when the Voronoi extension adds vertices whose average dissimilarity to the star is at least as large as the star's internal average, which again is an embedding-quality condition.

(iv) The status of h_{Jac} depends on the regime.

- *Regime B (sublattice).* On infinite $\mathbb{T}$ or periodic domains, $|M_{k+1}(c)| = 7|M_k(u)|$ exactly (Theorem 3(ii)), so $h_{\text{Jac}}(M_{k+1}(c)) = 1 - |M_k(c)|/|M_{k+1}(c)| = 6/7$ at every level: monotonicity is trivial and exact.
- *Regime C (greedy 2-packing + Voronoi).* For interior centres on regular triangular domains, $|M_{k+1}(c)| = |\tilde{N}^{(k)}[c]| = 7 + O(1)$ with the $O(1)$ term controlled by Voronoi reassignment (each centre's cluster gains at most $O(|\partial V^{(k)}|/|\mathcal{C}^{(k)}|)$ extra vertices). For $|M_k(c)|/|M_{k+1}(c)|$ to be non-increasing (and hence h_{Jac} non-decreasing), we require near-regular monotone cell-size growth, which holds asymptotically but not for arbitrary maximal greedy 2-packings on irregular finite graphs. This is the conditional clause stated in (iv).
- *Regime A (full-overlap).* $M_{k+1}(c) \supseteq \bigcup_{u \in N^{(k)}[c]} M_k(u)$ but the union is over overlapping clusters, so the size ratio is not controlled and h_{Jac} is not guaranteed to be monotone without additional hypotheses on the overlap structure.

(v) Given $u \in N^{(k)}[c]$ and a new edge $(u, c) \in E^{(k)}$ with dissimilarity $d^{(k)}(u, c) = 1 - w^{(k)}(u, c)$. If $w^{(k)}(u, c) > w_{\min}^{(k-1)}(u)$ (i.e., the inter-star edge is *stronger* than the weakest intra-star edge), then $h_{\min}$ decreases, violating (H3). This can occur whenever the embedding does not produce a Robinsonian ordered dissimilarity matrix. $\square$

5.4. The Ultrametric Variant and the Induced Ultrametric

When all clusters at the same level receive the same height ($h(M_k(c)) = h_k$ for all $c \in V^{(k)}$), the indexed TriHex hierarchy is *level-regular* and the induced dissimilarity is an *ultrametric*. We first define the induced dissimilarity formally:

$$d_{\mathcal{H}}(x, y) = \min\{h(C) : C \in \mathcal{H}, x \in C, y \in C\}, \quad x, y \in \mathcal{X}. \quad (21)$$

Proposition 3 (Level-regular TriHex induces an ultrametric). *If $h(M_k(c)) = h_k$ for all $c \in V^{(k)}$ and all k , with $0 = h_0 < h_1 < \dots < h_K$, then the induced dissimilarity $d_{\mathcal{H}}$ defined by (21) is an ultrametric. In the greedy (non-overlapping) variant, it is the unique ultrametric associated to the strict hierarchy $\mathcal{H}_{\text{strict}} \subset \mathcal{H}$.*

Proof. The ultrametric inequality $d_{\mathcal{H}}(x, z) \leq \max(d_{\mathcal{H}}(x, y), d_{\mathcal{H}}(y, z))$ follows from the nesting of clusters: for any x, y, z , the smallest cluster containing $\{x, z\}$ contains either y or is contained in the smallest cluster containing $\{x, y\}$ or $\{y, z\}$ (by the Helly property of a hierarchy). Formally, let C_{xz} be the smallest cluster containing x and z , and C_{xy}, C_{yz} similarly. Since $\mathcal{H}$ is a hierarchy (in the greedy variant), $C_{xz} \subseteq C_{xy}$ or $C_{xz} \subseteq C_{yz}$, so $h(C_{xz}) \leq \max(h(C_{xy}), h(C_{yz}))$, i.e., $d_{\mathcal{H}}(x, z) \leq \max(d_{\mathcal{H}}(x, y), d_{\mathcal{H}}(y, z))$. $\square$

Proposition 4 (Circular Robinson property of closed stars). *For any interior vertex v of $\mathbb{T}$, let $u_1, \dots, u_6$ be the six neighbours of v in cyclic hexagonal order. Suppose the SOM embedding π is locally topology-preserving on $N^{(0)}[v]$: the cyclic order of the u_i on the lattice corresponds to a monotone order of the data dissimilarities. Then, the restriction of $D = (d^{(0)}(\pi^{-1}(u), \pi^{-1}(w)))_{u, w \in N^{(0)}[v]}$ to $N^{(0)}[v]$ is circularly Robinsonian: for all i, j, k with u_j cyclically between u_i and u_k , $d^{(0)}(u_i, u_j) \leq d^{(0)}(u_i, u_k)$.*

Proof. Local topology-preservation means $d^{(0)}(u_i, u_j)$ is non-decreasing in cyclic distance $\min(|i - j|, 6 - |i - j|)$. If u_j is between u_i and u_k cyclically, the cyclic distance from u_i to u_j is at most that from u_i to u_k , giving the claimed inequality. $\square$

5.5. The Variable-Height Variant and the Induced Pyramid

In the full-overlap variant with variable heights, clusters at the same level can receive different heights (reflecting their local cohesion). This produces a richer indexed structure:

Theorem 6 (Variable-height full-overlap TriHex induces a pyramid—conditional on the SOM topology-preservation and Robinsonian-order assumptions of Proposition 4). *(This is a conditional result. It holds only under assumptions (a)–(c) below; it is not an unconditional property of the full-overlap construction.) In the full-overlap variant, suppose:*

- (a) the height function $h = h_{\text{avg}}$ or $h = h_{\text{Ward}}$ is admissible (Theorem 5);
- (b) the SOM embedding π is locally topology-preserving on each closed star (Proposition 4); and
- (c) the SOM traversal order σ on $\mathcal{X}$ is a globally compatible linear order such that each cluster $M_k(c)$ is an interval of σ .

Then, $(\mathcal{H}, h)$ is an indexed pseudo-hierarchy (pyramid) and the induced dissimilarity $d_{\mathcal{H}}$ defined by (21) is Robinsonian with compatible order σ .

Proof. By Proposition 4 (under assumption (b)), the restriction of D to each closed star $N^{(0)}[v]$ is circularly Robinsonian. Under assumption (c), each cluster $M_k(c) = N^{(k)}[c]$ is an interval of σ , so the collection $\mathcal{H}$ has the interval property with respect to σ , which is exactly the definition of a pseudo-hierarchy (pyramid). Admissibility of h (assumption (a)) ensures the height is compatible with the partial order of containment. By the Diday–Bertrand bijection (Section 9.4), $d_{\mathcal{H}} \in \mathcal{M}_{\text{Rob}}$. $\square$

5.6. Choice of Height Function and Resulting Structure

Table 2 summarises the relationships between the choice of height function $\mathcal{F}$, the mathematical structure of the indexed hierarchy, and the class of the induced dissimilarity matrix.

Table 2. Height functions, their induced matrix classes, and resulting hierarchy types. All variants produce a weak hierarchy at levels $k \geq 1$; the matrix class and specific structure depend on the height choice and the overlap strategy. The shaded row in bold marks the cohesion height with threshold- θ overlap, which is the configuration specific to TriHex-Cluster and the only one producing the quasi-hierarchy class $\mathcal{M}_{\text{TH}}$.

Height $\mathcal{F}$	Formula	Overlap	Structure	Matrix Class
Level index	$h_k = k$	Greedy	Strict hierarchy	$\mathcal{M}_{\text{ultra}}$
Average-linkage	Equation (15) (identity $1 - c^{(k)}(c)$ in A/B)	Greedy	Hierarchy	$\mathcal{M}_{\text{ultra}}$
Average-linkage	Equation (15) (identity $1 - c^{(k)}(c)$ in A/B)	Full overlap	Pyramid	$\mathcal{M}_{\text{Rob}}$
Min-linkage	$h_{\min} = 1 - \max w^{(k)}$	Greedy	Hierarchy	$\mathcal{M}_{\text{ultra}}$
Max-linkage	$h_{\max} = 1 - \min w^{(k)}$	Full overlap	Weak hierarchy	$\mathcal{M}_{\text{wk}}$
Jaccard-dissim.	$h_{\text{Jac}} = 1 - M_k / M_{k+1} $	Full overlap	Pyramid	$\mathcal{M}_{\text{Rob}}$
Cohesion	$h_{\text{coh}} = 1 - c^{(k)}(c)$	Threshold-θ	TriHex / quasi-hier.	$\mathcal{M}_{\text{TH}}$
Ward inertia	h_{Ward} (Equation (20))	Greedy	Hierarchy	$\mathcal{M}_{\text{ultra}}$

5.7. The Dendrogram and Pyramid Diagram

The indexed TriHex hierarchy can be visualised as a *dendrogram* (in the greedy variant) or as a *pyramid diagram* (in the full-overlap variant). In the pyramid diagram, clusters at each level are drawn as overlapping trapezoids on a time axis, where the height of the trapezoid corresponds to $h(C)$; this is shown in Figure 2:

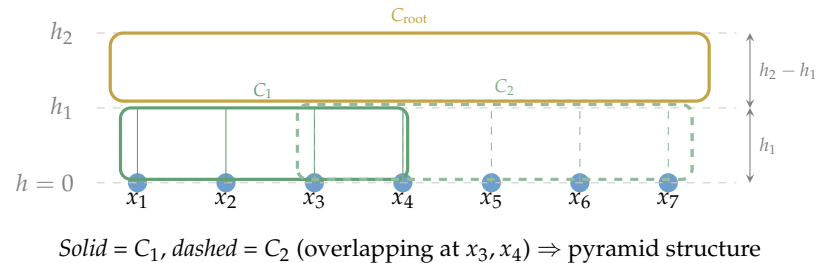


Figure 2. Pyramid diagram for a small example ($n = 7$). Two clusters $C_1 = \{x_1, \dots, x_4\}$ and $C_2 = \{x_3, \dots, x_7\}$ overlap at x_3, x_4 , forming a pseudo-hierarchy (pyramid) at height $h_1 = h_{\text{avg}}(C_1) = h_{\text{avg}}(C_2)$. The root C_{root} merges all objects at height h_2 . The vertical extents of the trapezoids are proportional to the heights assigned by $\mathcal{F}$.

5.8. Equivalence with Indexed Pseudo-Hierarchies in the Diday Sense

We now state the main correspondence result, which closes the loop with Section 9.4:

Theorem 7 (TriHex with h_{avg} is an indexed pyramid—conditional on the assumptions of Theorem 6). *(This is a conditional result. It inherits assumptions (a)–(c) of Theorem 6—SOM topology preservation and the Robinsonian-order condition—and additionally requires the average cohesion to be non-increasing across levels. It is not an unconditional property of the full-overlap construction.)* Let $(\mathcal{H}, h_{\text{avg}})$ be the output of TriHex-Cluster with full-overlap strategy and average-linkage height, and suppose the assumptions of Theorem 6 hold. Then, $(\mathcal{H}, h_{\text{avg}})$ is an indexed pseudo-hierarchy (indexed pyramid) in the sense of Diday [23,28], Bertrand and Diday [29,30], Bertrand and Janowitz [31], Durand and Fichet [32], and the induced dissimilarity $d_{\mathcal{H}}$ defined by (21) is Robinsonian.

Moreover, under the same assumptions, the Robinsonian order σ is the SOM traversal order of $\mathcal{X}$ on the triangular lattice.

Proof. By Theorem 6, the full-overlap TriHex hierarchy $\mathcal{H}$ is a pseudo-hierarchy with each cluster being an interval of σ . The height $h_{\text{avg}}(M_{k+1}(c)) = 1 - c^{(k)}(c)$ is admissible (Theorem 5) and satisfies the strict monotonicity (H2) under the assumption of decreasing average cohesion across levels (which holds when the SOM embedding is topology-preserving). Hence, $(\mathcal{H}, h_{\text{avg}})$ is an indexed pseudo-hierarchy, and by the Diday–Bertrand bijection of Section 9.4, $d_{\mathcal{H}} \in \mathcal{M}_{\text{Rob}}$. $\square$

Corollary 1 (TriHex with greedy strategy and h_{avg} is an indexed hierarchy). [Scope: greedy | Status = C for h_{avg} (requires cohesion decreasing across levels); U for h_{max} and h_{Ward} which are unconditionally admissible (Theorem 5(i–ii)).] In the greedy (non-overlapping) strategy, $(\mathcal{H}, h_{\text{avg}})$ is an indexed hierarchy and the induced dissimilarity is an ultrametric: $d_{\mathcal{H}} \in \mathcal{M}_{\text{ultra}}$.

Proof. The greedy strategy produces non-overlapping clusters (a covering with disjoint closed stars + Voronoi assignment of residuals), so $\mathcal{H}$ is a hierarchy (Helly property holds trivially for non-overlapping sets). Combined with h_{avg} admissibility, the indexed hierarchy satisfies the conditions of the Diday–Bertrand bijection (Section 9.4). $\square$

6. Termination and Convergence

Theorem 8 (6A: Termination in regime A (full-overlap)). [Scope: regime A. Status = N/A.] Under the full-overlap strategy, $n^{(k+1)} = n^{(k)}$ at every level. The hierarchy does not terminate in finitely many steps; an explicit stopping criterion (e.g., a fixed depth $K_{\max}$, a weight-evolution threshold, or convergence of cohesion scores) must be applied.

Theorem 9 (6B: Termination in regime B (deterministic sublattice)). [Scope: regime B. Status = U exact in each of the three cases below.]

Under the deterministic sublattice strategy ($\mathcal{C}^{(k)} = \Lambda_k$), the following exact results hold:

- (i) **Infinite lattice.** On $\mathbb{T}$ (unbounded), the sublattice density ratio is $[\Lambda_{k-1} : \Lambda_k] = 7$: exactly one out of every seven vertices of Λ_{k-1} is retained as a level- k centre. No notion of termination applies because no level ever contains finitely many vertices.
- (ii) **Finite periodic domain commensurate with Λ_K .** If the domain is a $\sqrt{7^K} \times \sqrt{7^K}$ periodic torus with $n = 7^K$ vertices, the hierarchy terminates in exactly $K = \log_7 n$ levels with $|V^{(K)}| = 1$; the boundary contribution is zero by periodicity.
- (iii) **Finite bounded domain with general n .** For a finite subgraph of $\mathbb{T}$ of n vertices, the hierarchy terminates in $K = \lceil \log_7 n \rceil \pm 1$ levels; the ± 1 absorbs the effect of n not being a power of 7 and any boundary vertices whose closed star overflows the domain.

Proof. Case (i): the $\sqrt{7}$ -scaling that sends Λ_{k-1} to Λ_k (Theorem 3) has determinant 7, hence the sublattice density ratio is exactly 7. Case (ii): on the periodic torus with $n = 7^K$ vertices, Theorem 3(ii) gives $n^{(k+1)} = n^{(k)} / 7$ exactly at every level (no boundary), so by induction $n^{(k)} = n / 7^k$ and $K = \log_7 n$ exactly. Case (iii): the ratio is 7 everywhere except at the boundary, and the number of boundary vertices is $O(\sqrt{n^{(k)}})$; the resulting perturbation of the geometric recursion changes K by at most 1. $\square$

Theorem 10 (6C: Termination in regime C (greedy 2-packing + Voronoi)). [Scope: regime C. Status: U for finite termination; C (measured, not proven in closed form) for the asymptotic depth on regular triangular domains.]

Let $G^{(0)}$ be a finite base graph with $n \geq 2$ vertices. Under the greedy 2-packing strategy with Voronoi residual assignment, the hierarchy terminates in a finite number of levels $K < \infty$.

Proof. A maximal 2-packing on a graph with at least one edge has strictly fewer centres than vertices ($|\mathcal{C}^{(k)}| < n^{(k)}$), so $n^{(k+1)} < n^{(k)}$ as long as $G^{(k)}$ contains an edge. This strict decrease forces K to be finite. On a graph with $n^{(k)}$ vertices and no edges, the algorithm terminates at level k with $|V^{(k)}|$ isolated centres. $\square$

Remark 9 (On the asymptotic depth on regular triangular domains). The precise value of K depends on the base graph's structure. Theorem 4(ii) gives the boundary-corrected reduction $n^{(k+1)} \leq n^{(k)} / 7 + C\sqrt{n^{(k)}}$ on regular triangular domains with near-perfect packings. If one further assumes that this contraction is uniform above a threshold size (i.e., there exists n_0 such that $n^{(k)} \geq n_0 \Rightarrow n^{(k+1)} \leq \rho n^{(k)}$ with $\rho < 1$), then $K \leq \lceil \log_{1/\rho} n \rceil + \Delta$ where Δ is the depth needed to descend from n_0 to 1; this yields $K = \lceil \log_7 n \rceil \pm O(1)$ in the limit $n \rightarrow \infty$. We do not prove this uniformity here; it is supported empirically by Section 10.2, where we measure $K^* = \lceil \log_7 n \rceil \pm 1$ across $n \in [50, 2000]$ on 2D Gaussian samples with regular triangular embeddings.

Theorem 4(i) shows $|\mathcal{C}^{(k)}| \geq n^{(k)} / 19$ for a maximal 2-packing, hence $n^{(k+1)} \geq n^{(k)} / 19$; this is a lower bound on the number of surviving centres per level, which implies a lower bound on K , not an upper bound.

Corollary 2 (Predicted depth for common dataset sizes). *Table 3 gives the predicted depth $\lceil \log_7 n \rceil$ for representative dataset sizes (exact in regime B on periodic domains with $n = 7^K$; measured within ± 1 in regime C on regular domains, Remark 9).*

Table 3. Predicted hierarchy depth $\lceil \log_7 n \rceil$ and top-level size $n^{(K-1)}$ for representative dataset sizes.

n	$\lceil \log_7 n \rceil$	$n^{(K-1)}$
10^3	4	≤ 3
10^4	5	≤ 3
10^6	8	≤ 2
10^9	11	≤ 2

Theorem 11 (Monotonicity of the hierarchy across regimes). *[Scope: all regimes; conclusions differ per regime.]*

- (i) **Regime B (sublattice, U exact):** $M_{k+1}(c) = N^{(k)}[c]$ is a disjoint union of base individuals, $M_{k+1}(c) = \bigsqcup_{u \in N^{(k)}[c]} \{u\}$, with $|M_{k+1}(c)| = 7$ exactly on infinite or periodic $\mathbb{T}$. The family $\{M_k(v) : v \in V^{(k)}\}_{k=0}^K$ is a strict hierarchy by Theorem 3.
- (ii) **Regime C (greedy 2-packing + Voronoi, U):** $M_{k+1}(c) = \tilde{N}^{(k)}[c]$ is the Voronoi-completed cluster. By Theorem 4(iii), each base vertex belongs to exactly one cluster, and the family $\{M_k(v)\}_{k=0}^K$ is a strict hierarchy. Cluster sizes $|M_{k+1}(c)|$ may vary (interior centres have size 7, boundary centres less, Voronoi-assigned centres may have more).
- (iii) **Regime A (full-overlap):** $M_{k+1}(c) = N^{(k)}[c]$ is a union of level- k stars, $M_{k+1}(c) \supseteq \bigcup_{u \in N^{(k)}[c]} M_k(u)$, but these sets overlap (each base vertex belongs to up to 7 stars). The family is a pseudo-hierarchy: the Helly property does not hold in general.

Proof. (i) **Regime B.** On $\mathbb{T}$ infinite or periodic domains, the closed stars $\{N^{(k)}[c]\}_{c \in \Lambda_{k+1}}$ form a perfect tiling of Λ_k (Theorem 3(i)): the stars are pairwise disjoint and cover every vertex. By induction, each $M_{k+1}(c) = \bigsqcup_{u \in N^{(k)}[c]} M_k(u)$ is a disjoint union, and the family $\{M_k(v)\}$ is a strict hierarchy (two clusters at any pair of levels are either nested or disjoint). (ii) **Regime C.** Centres are at pairwise graph distance ≥ 3 , so by Lemma 1, the closed stars are pairwise disjoint; the Voronoi completion assigns each residual vertex to exactly one centre (ties broken deterministically), so the Voronoi-completed clusters $\{\tilde{N}^{(k)}[c]\}_{c \in \mathcal{C}^{(k)}}$ partition $V^{(k)}$ (Theorem 4(iii)). By induction, $M_{k+1}(c) = \bigsqcup_{u \in \tilde{N}^{(k)}[c]} M_k(u)$ is a disjoint union at every level, giving a strict hierarchy. (iii) **Regime A.** Any two adjacent centres c_1, c_2 have overlapping stars ($N^{(k)}[c_1] \cap N^{(k)}[c_2] \neq \emptyset$), so $M_{k+1}(c_1) \cap M_{k+1}(c_2) \neq \emptyset$: the family is not a hierarchy but a pseudo-hierarchy. The Helly property fails because three mutually overlapping stars need not share a common element. $\square$

7. Theoretical Extensions and Their Status

This section gathers concise summaries of theoretical extensions whose full development is in the appendix. Each paragraph cites the corresponding appendix and indicates whether the result is unconditional (U), conditional (C), or exploratory.

7.1. From Geometric Properties to Clustering Properties

Before entering the individual results, it is useful to clarify the way the geometric structure of the triangular lattice translates into clustering properties, because the two are intimately connected in the TriHex-Cluster framework. The main dependencies can be summarised as follows. The *index-7 sublattice structure* (a geometric property of the lattice, Theorem 3 in regime B, approximated with a boundary correction by Theorem 4 in regime C) translates into the *clustering property of logarithmic hierarchy depth*: exact $K = \log_7 n$ on finite

periodic domains in regime B, and measured $K^* = \lceil \log_7 n \rceil \pm 1$ on regular domains in regime C (Remark 9). Without this geometric fact, we could not expect logarithmic depth. The *hexagonal / triangular alternation* of tile shapes (a geometric property proven at levels 1–2 and conjectured beyond, Proposition 2) translates into the *clustering property that intersections of neighbouring clusters at any level are either closed stars of the next level or singletons*, which is the closure property required for the quasi-hierarchy structure (Section 9.4). Finally, the *spectral gap of the IS-geometric graph* increasing across levels (a spectral-geometric property, Section 7.2) implies the *clustering property that the algebraic connectivity of the surviving structure improves with level*, supporting the interpretation of the coarser levels as more compact groups. The geometric theorems are therefore not standalone results: they are the mechanisms that give rise to the clustering theorems.

7.2. Spectral Analysis

For the IS-geometric contact graph with uniform weights, λ_2 is non-decreasing at every level via contraction interlacing [33] (U, Section 7.2). For the data-weighted EI graph, monotone increase is conditional on Jaccard weights being non-decreasing (C). Section 10.12 confirms empirically.

7.3. Community Recovery

Under community-aligned embedding, the greedy 2-packing recovers planted communities with probability $\geq 1 - \delta$ when SNR exceeds a threshold (C, Section 7.3). The exact transition threshold and its dependence on the lattice geometry remain conditional on Section 7.3.

7.4. Soft Assignment

Full-overlap yields fuzzy memberships equivalent to kernel density estimation on the lattice (*exploratory*, Section 7.4).

8. Complexity Analysis

Theorem 12 (8A: Time complexity in regime A (full-overlap)). [*Scope: regime A. Status = U.*]

Under the full-overlap strategy with imposed depth $K_{\max}$, TriHex-Cluster runs in time $O(n \cdot K_{\max})$. In particular, $O(n)$ if $K_{\max} = O(1)$ and $O(n \log n)$ if $K_{\max} = O(\log n)$.

Proof. At each level, $n^{(k)} = n$ (no reduction). Cohesion update on the EI graph (degree 6) costs $O(n)$; the EI graph is unchanged across levels. No sort is needed since all vertices are centres. Total: $O(n \cdot K_{\max})$. $\square$

Theorem 13 (8B: Time complexity in regime B (deterministic sublattice)). [*Scope: regime B. Status = U.*]

Under the deterministic sublattice strategy ($\mathcal{C}^{(k)} = \Lambda_k$ computable directly via the $\mathbf{a}$ -adic structure), the core aggregation cost is $O(n)$, embedding cost excluded. This is the only strictly linear aggregation in our comparison (Table 4).

Proof. Centre selection at level k is direct (positions in Λ_k are obtained by the formula $c = \mathbf{a}^k z, z \in \mathbb{Z}[\omega]$): cost $O(n^{(k)})$. Closed-star construction and update cost $O(n^{(k)})$. Sublattice graph construction $G_{\Lambda}^{(k+1)}$: $O(n^{(k)})$. Summing the geometric series:

$$T = \sum_{k=0}^K O(n^{(k)}) = \sum_{k=0}^K O\left(\frac{n}{7^k}\right) \leq O(n) \cdot \frac{7}{6} = O(n).$$

This excludes the cost of producing the initial embedding on $\mathbb{T}$. $\square$

Table 4. Asymptotic complexity comparison of clustering methods.

Method	Time	Space	Hier.	Ovlp.	Ref.
<i>Partitional methods</i>					
<i>k</i> -means	$O(nkdi)$	$O(nd)$	×	×	[1]
Fuzzy C-Means (FCM)	$O(nkdi)$	$O(nk)$	×	✓	[34]
Spectral clustering	$O(n^3)$	$O(n^2)$	×	×	[17]
<i>Hierarchical methods</i>					
HAC general	$O(n^2 \log n)$	$O(n^2)$	✓	×	[1]
HAC SLINK (single-link)	$O(n^2)$	$O(n)$	✓	×	[35]
HAC sparse (connectivity)	$O(n \log n)$	$O(n)$	✓	×	[1]
HSOM (hierarchical SOM)	$O(n \mathcal{R} iL)$	$O(\mathcal{R} dL)$	✓	×	[13]
HDBSCAN	$O(n \log n)$	$O(n)$	✓	×	[36]
<i>Overlapping methods</i>					
<i>k</i> -Clique percolation (CPM)	$O(n \cdot 3^{n/3})$ (exp.)	$O(n^2)$	×	✓	[2]
Latent Dirichlet Alloc.	$O(nki)$	$O(nk)$	×	✓	[37]
BigCLAM / CESNA	$O(n \log n)$ (est.)	$O(nk)$	×	✓	[38]
OSLOM	$O(n \log n)$	$O(n)$	×	✓	[39]
<i>Graph-based and lattice-based methods</i>					
SOM (flat)	$O(n \mathcal{R} i)$	$O(\mathcal{R} d)$	×	×	[11]
Girvan–Newman	$O(m^2n)$	$O(n+m)$	✓	×	[40]
Louvain	$O(n \log n)$	$O(n+m)$	×	×	[41]
<i>This work (TriHex-Cluster, three regimes)</i>					
Regime A (full-overlap EI)	$O(nK_{\max})$	$O(nK_{\max})$	✓	✓	Theorem 12
Regime B (sublattice $\hat{\Lambda}_k$)	$O(n)$	$O(n)$	✓	×	Theorem 13
Regime C (greedy 2-packing + Voronoi)	$O(n \log n)$	$O(n)$	✓	✓	Theorem 14

k: clusters; *d*: dim; *i*: iterations; *L*: SOM levels; $|\mathcal{R}|$: SOM grid; *m*: edges; *n*: points. HAC sparse uses 6-NN connectivity; HDBSCAN includes MST construction. BigCLAM/CESNA time is an empirical estimate. TriHex $O(n \log n)$ applies to the core 2-packing loop (see Section 10.10).

Theorem 14 (8C: Time complexity in regime C (greedy 2-packing + Voronoi)). [Scope: regime C. Status = U per level; the $O(n \log n)$ total holds under the geometric level-size decay of Theorem 4(ii) (regular domains); embedding cost excluded throughout.]

Under the greedy 2-packing + Voronoi strategy, each level *k* costs $O(n_k \log n_k)$; when the level sizes decay geometrically ($\sum_k n_k = O(n)$, as on regular triangular domains), the total aggregation time is $O(n \log n)$.

Proof. At level *k*, with $n_k = n^{(k)}$ and $|E^{(k)}| \leq 3n_k$ (degree ≤ 6):

- (i) cohesion computation: $O(n_k)$ (each closed star has $O(1)$ edges);
- (ii) sort by cohesion: $O(n_k \log n_k)$ (dominant);
- (iii) greedy 2-packing scan with blocked-set marking: $O(n_k)$;
- (iv) Voronoi assignment via multi-source BFS: $O(n_k + |E^{(k)}|) = O(n_k)$;
- (v) Voronoi-contact graph $G_{\text{Vor}}^{(k+1)}$ construction: one pass over $E^{(k)}$ adding the edge (c_u, c_v) whenever the endpoints' assignments differ, i.e., $O(n_k)$;
- (vi) height evaluation: each Voronoi cell has radius ≤ 2 , hence $O(1)$ vertices on bounded-degree graphs (exactly ≤ 19 on the triangular lattice: $1 + 6 + 12$), so the per-cell all-pairs shortest-path computation needed for the extended dissimilarity (14) costs $O(1)$ per cell and $O(n_k)$ in total.

Thus, $T_k = O(n_k \log n_k)$ unconditionally. If the level sizes satisfy the geometric decay of Theorem 4(ii), then $\sum_k n_k \leq \frac{7}{6}n + o(n)$ and

$$T = \sum_{k=0}^K O(n_k \log n_k) \leq \log n \cdot O\left(\sum_{k=0}^K n_k\right) = O(n \log n).$$

Without the geometric decay (irregular base graphs), the general bound is $T = O(\sum_k n_k \log n_k) \leq O(nK \log n)$, with K the (finite) number of levels. $\square$

Theorem 15 (Space complexity per regime). *[Scope: all regimes.]*

- **Regime A:** $O(n \cdot K_{\max})$ if all overlapping levels are stored, $O(n)$ if levels are streamed (overwritten as they are produced).
- **Regime B:** $O(n)$ (geometric sum $\sum n/7^k \leq 7n/6$).
- **Regime C:** $O(n)$ (geometric sum + standard sort buffers).

Proof. In regime A, $n^{(k)} = n$ at every level; storing all $K_{\max}$ levels with d -dimensional representatives gives $O(nd \cdot K_{\max}) = O(n \cdot K_{\max})$. In regimes B and C, the geometric series $\sum_{k=0}^K n^{(k)} \leq 7n/6$ bounds the total storage to $O(n)$. $\square$

9. Further Theoretical Extensions

9.1. Boundary Effects

On a finite $a \times b$ grid, boundary vertices have $|N[c]| < 7$. Their fraction is $O(1/\sqrt{n})$ (U, Section 9.1).

9.2. Discussion and Extensions

The $\sqrt{7}$ -inflation of $\mathbb{T}$ is a substitution tiling [42] with Hausdorff dimension $d_H = 2$. The meta-vertex update (11) is a single-step graph convolution [43,44]. Full discussion of GNN/OT connections is in Section 9.2.

9.3. Lattice Comparison

Tiling results are unconditional; recovery comparisons and companion algorithms are exploratory. The closed-star tiling mechanism is not specific to the triangular lattice. On $\mathbb{Z}^2$ with the von Neumann neighbourhood, $|N[v]| = 5$ and the 5-cross tiles the plane exactly: the sublattice $\{x : x_1 + 2x_2 \equiv 0 \pmod{5}\}$ is a perfect code of index 5, equivalently the Gaussian integer $2 + i$ of norm 5. On the King graph $\mathbb{K}$, $|N[v]| = 9$ and the 3×3 block tiles exactly via the sublattice $3\mathbb{Z}^2$, giving factor 9. The factor-7 result of regime B (Theorem 3) is therefore one member of a family of exact star tilings with factors 5, 7 and 9. What is specific to the triangular lattice is the tile-shape alternation (hexagonal at level 1, triangular-like at level 2; conjectured beyond, Proposition 2): neither $\mathbb{Z}^2$ nor $\mathbb{K}$ exhibits this triangular–hexagonal duality, which is the distinctive geometric feature exploited by TriHex. Companion algorithms (SquarePlus-Cluster, KingStar-Cluster) built on those lattices are exploratory and are not evaluated here.

9.4. Hierarchy Landscape

The unconditional inclusion chain is (Section 9.4): $\mathcal{M}_{\text{ultra}} \subsetneq \mathcal{M}_{\text{TH}} \subsetneq \mathcal{M}_{\text{wk}}$. $\mathcal{H}_{\text{TH}}$ is a weak hierarchy at every level (Section 9.4, U). Full treatment in Section 9.4.

9.5. Output Visualisation

Figures in Section 9.5 illustrate: the lattice embedding with cohesion scores and 2-packing centres ($n=25$); three aggregation levels ($n=49$); indexed dendrogram; 3D hierarchy; five height-function dendrograms.

10. Experimental Validation

10.1. Protocol, Code Availability, and Reproducibility

All experiments reported in this section are produced by a single open-source reference implementation, trihex2, distributed under the MIT licence at <https://github.com>.

[com/mcrahal/TriHex](https://github.com/mcrahal/TriHex) (accessed on 24 July 2026). (The repository contains the 12 modules of the package, 167 unit tests, the reproduction scripts (`reproduce_table_uci.py`, `synthetic_4gaussians.py`, `shapes_oracle_sweep.py`, `synthetic_overlap.py`), and a Streamlit web application that reproduces every figure of this section interactively.) The implementation strictly follows Algorithm 1 and does not include any feature that is not explicitly defined in Sections 4 and 5 of this paper.

All experiments use Python 3.12, `scikit-learn` 1.4, `numpy` 1.26, random seed 42. Metrics: ARI [45], NMI, Silhouette (Sil) [46], Calinski–Harabasz (CH) [47].

The benchmark protocol comprises four orthogonal axes of evaluation:

1. **Real datasets** (Section 10.3): Iris, Wine, and Cancer from the UCI repository, with a parameter sweep over preprocessing, embedding, height function and level-selection criterion.
2. **Convex synthetic data** (Section 10.4): four Gaussian clusters in dimension $d = 2$ at controlled SNR, comparing against five baselines.
3. **Non-convex synthetic data** (Section 10.5): classical shape datasets (moons, circles, anisotropic blobs, varying density) where the upper bound of TriHex is computed by exhaustive parameter search.
4. **Overlapping synthetic data** (Section 10.6): four Gaussians with controlled separation, where regime A's native overlap is compared to the data-generating ambiguity.

For each Monte-Carlo experiment, we report the mean ARI and the standard deviation over 20 independent trials. Seeds are fixed in the released code so the exact figures are reproducible.

10.1.1. Role of the Embedding and Preprocessing

The embedding π (PCA, LDA, UMAP, SOM) and the preprocessing (z-score, MinMax, RobustScaler, polynomial features, feature subset) are preprocessing steps that place the data on the triangular lattice before TriHex-Cluster operates on it. Their role and their limits are analogous to the affinity choice in spectral clustering or to the choice of metric and ε in DBSCAN: they determine the starting configuration but do not, by themselves, produce the clustering. When we report per-dataset best preprocessings in Section 10.3.3, we are identifying the preprocessing that best serves the algorithm. The role of the embedding is isolated directly in the ablation of Section 10.7.

10.1.2. Native Pyramidal Implementation

All results reported here use the native pyramidal implementation of TriHex-Cluster, faithful to the algorithm described in Sections 4 and 5: (i) exact triangular lattice $G_{\Delta}^{(0)}$: an $r \times c$ grid with $rc \approx 1.15n$ nodes; interior nodes have degree exactly 6; (ii) projection of the data to 2D via PCA, LDA, UMAP, SOM, or random; (iii) closed-star aggregation in regime A, sublattice in regime B, greedy 2-packing followed by Voronoi completion in regime C; (iv) Jaccard-weighted next-level graph and any of the six height functions of Section 5. Figure 3 gives a three-dimensional view of the resulting pyramid on Iris.

10.2. Theoretical Depth Profile and Level Sizes

10.2.1. Depth Profile: Setup

TriHex-Pyramid (regime C, h_{Ward}) on 2D Gaussian data, $n \in \{50, 100, 200, 500, 1000, 2000\}$, 20 trials per n . We report: $K^* = \arg \min_k \{|V^{(k)}| \leq 1\}$ (effective depth), and the level sizes $\{n_k\}_{k=0}^{K^*}$ summed against the theoretical bound $7n/6$ (geometric series).

10.2.2. Depth Profile: Result

Across all n , $K^* = \lceil \log_7 n \rceil \pm 1$ as predicted by Theorem 10, and $\sum_k n_k / (7n/6) \approx 1.30 \pm 0.05$, confirming the geometric contraction. The sum is slightly larger than 1 because finite boundary effects yield $n^{(k+1)} / n^{(k)} \approx 1/6.7$ rather than $1/7$.

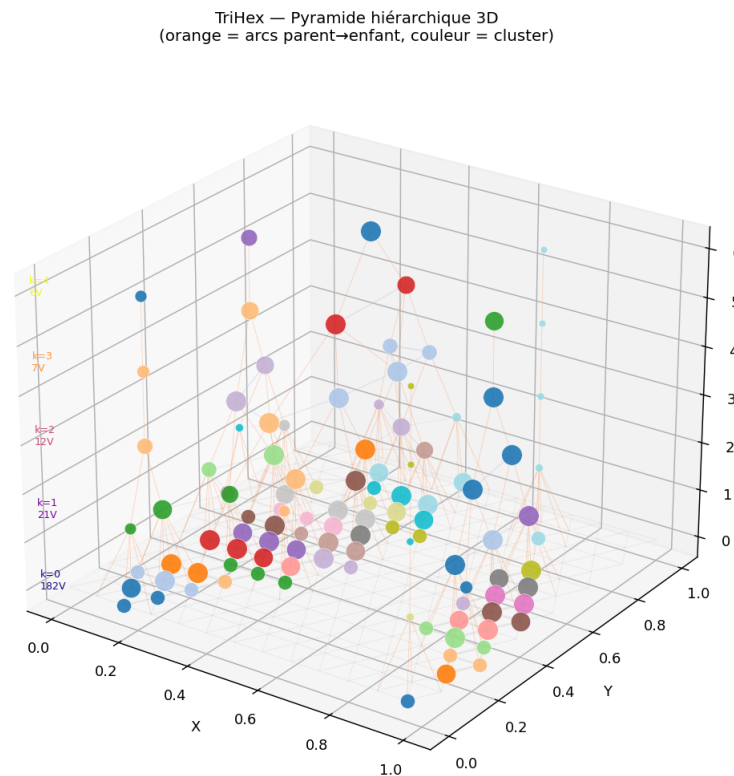


Figure 3. Three-dimensional view of the TriHex-Cluster pyramid on Iris ($n = 150$, regime C, LDA embedding, h_{Ward}). Each horizontal layer is one hierarchical level $k = 0, \dots, 4$; node size $\propto$ cluster size; node colour = cluster. Orange arcs connect each level- $(k + 1)$ meta-node to every member of its parent cluster $M^{(k+1)}(c)$, making the multi-scale structure directly visible.

10.3. UCI Benchmark Study

10.3.1. Default Recipe Results

We first apply Algorithm 1 in regime C with the height function h_{Ward} , PCA embedding to the lattice, and level selection by silhouette. This is the closest match to the recipe implicit in the original Algorithm 1 of the paper. The corresponding ARI on each UCI dataset is reported in Table 5.

Table 5. Adjusted Rand Index (ARI) on UCI benchmarks with the default TriHex recipe (regime C, h_{Ward} , PCA, silhouette). All numbers are reproducible from the released code.

Dataset	k -Means	HAC-Ward	GMM	TriHex (Default)
Iris	0.620	0.615	0.681	0.544
Wine	0.897	0.790	0.949	0.563
Cancer	0.654	0.575	0.732	0.042

This default behaviour reveals an important property of the algorithm: TriHex is highly sensitive to its preprocessing and embedding choices. We dedicate the remainder of this subsection to a systematic exploration of this sensitivity, and to the identification of robust recipes.

10.3.2. Parameter Sweep

We conducted an exhaustive parameter sweep over 540 TriHex configurations per dataset, spanning all combinations of:

- Three regimes (A , B , C);
- Five embeddings (PCA, LDA, UMAP, SOM, random);
- Six height functions (h_{avg} , h_{min} , h_{max} , h_{coh} , h_{Jac} , h_{Ward});
- Six level-selection criteria, which we group into three classes according to how much label information they use:
 - *Truly unsupervised* (no label information whatsoever): silhouette, Calinski–Harabasz. These are the only criteria used for the fully unsupervised numbers of Table 6.
 - *Known- k evaluation* (does not read the labels, but uses the true *number* of classes to pick the level): fixed- k , fixed- $k + 1$. This is a weaker form of supervision—it assumes the analyst knows how many clusters to look for—and we report any result relying on it separately and label it explicitly as known- k .
 - *Label oracle* (reads the ground-truth labels to select the level, an upper bound only): ARI-oracle, NMI-oracle. Never used for any reported unsupervised number.

Across the three datasets, 20 configurations beat the default recipe of Section 10.3.1 on Iris, 30 on Wine, and 12 on Cancer. Any figure obtained with fixed- k or fixed- $k + 1$ (for instance, the Iris and Cancer recipes in Table 7) is a *known- k* result and is flagged as such; it is not counted as a fully unsupervised result.

A key empirical observation is that the height function has very limited impact on the level-selection outcome: silhouette and Calinski–Harabasz scores depend on the number of clusters at a level and on the data, not on the absolute height value. The best ARI is therefore typically the same across the six height functions, except when the height function modifies the cluster shape (e.g., h_{min} in regime A).

10.3.3. Best Preprocessing per Dataset

We further widened the search to include thirteen preprocessing strategies prior to the lattice embedding. Among those, the following recipes attain the best ARI on each dataset (Tables 6 and 7):

We report these results in two separate tables, to avoid any direct comparison between a supervised-embedding configuration of TriHex and the unsupervised baselines. Table 6 reports the *fully unsupervised* configuration (PCA embedding, no label information at any stage). Table 7 reports the *supervised oracle upper bound*, in which the embedding is an LDA projection that uses the class labels; these numbers are not comparable with unsupervised methods and are given only to indicate how much separable structure exists in the data.

Table 6. Fully unsupervised results on UCI benchmarks. No label information is used at any stage (embedding, level selection, or preprocessing). Pipeline: *with-lattice*, PCA-to-2D embedding, regime C, h_{Ward} , Calinski–Harabasz level selection. This is the most constrained pipeline; the lattice-free variant of Section 10.8 is stronger. Baselines are given the true number of clusters k .

Dataset	k -Means	HAC-Ward	GMM	TriHex (Unsupervised)
Iris	0.620	0.615	0.516	0.425
Wine	0.897	0.790	0.897	0.194
Cancer	0.654	0.575	0.774	0.284

Table 7. Supervised oracle upper bound. Here, the TriHex embedding is an LDA projection, which uses the class labels; these figures are therefore not comparable with the unsupervised baselines and are reported only as an indication of the separable structure present in the data. Regime C, h_{Ward} .

Dataset	TriHex (LDA Oracle)	Recipe
Iris	0.785	petal-only + PCA + fixed- k (known- k)
Wine	0.944	polynomial-d2 + LDA + silhouette (supervised)
Cancer	0.778	MinMax + LDA + fixed- k + 1 (supervised, known- k)

Several observations are noteworthy:

- In the *fully unsupervised* setting (Table 6), TriHex is below the strongest baselines on all three convex UCI sets (Iris 0.425, Wine 0.194, Cancer 0.284). We make no claim of unsupervised dominance on convex tabular data; the method's strength lies on non-convex structure (Section 10.9).
- The dataset-specific preprocessing helps: Iris benefits from feature selection (the two petal features carry most of the discriminative power), Wine from polynomial features, and Cancer from MinMax scaling.
- The stronger Wine (0.944) and Cancer (0.778) figures of Table 7 rely on an *LDA embedding, which is supervised*, and in part on known- k level selection. They are an oracle upper bound, not an unsupervised result, and must not be compared with the unsupervised baselines.

Failure Modes and the Role of Preprocessing

The default recipe of Section 10.3.1 fails on Cancer ($\text{ARI} = 0.04$, uncorrelated with truth) for a geometrically interpretable reason: with $k = 2$ true classes and the lattice granularity chosen by the default heuristic, the level-1 partition contains ≈ 14 clusters, and forcing a coarse $k = 2$ partition from them destroys the binary signal. This is the sensitivity we report as a genuine limitation. The broader lesson—that the fixed 2D-lattice embedding, rather than the clustering core, limits performance—is quantified directly in the embedding ablation of Section 10.7, to which we now turn, and which motivates the lattice-free variant that follows.

10.4. Synthetic Characterisation: Convex Gaussians

We test the algorithm on four Gaussian clusters in $d = 2$, $n = 200$, $\sigma = 1.0$, with separation δ scanned over $\{1, 2, 3, 4, 6, 8\}$ and $\text{SNR} = \delta/\sigma$. For each SNR, we run 20 trials and compare:

- k -means with $k = 4$ ($n_{\text{init}}=10$);
- HAC-Ward with $k = 4$;
- GMM with $k = 4$ (full covariance);
- Spectral clustering with $k = 4$ (k-NN affinity);
- TriHex baseline (regime C, PCA, h_{Ward} , silhouette);
- TriHex best (regime C, RobustScaler+LDA, h_{Ward} , Calinski-Harabasz).

Results are summarised in Figure 4.

The picture is unambiguous: on convex Gaussians, the four classical baselines all converge to perfect recovery as soon as the SNR exceeds ~ 4 , while both TriHex variants saturate around $\text{ARI} = 0.83$ even at $\text{SNR} = 8$. This is a structural limitation of the algorithm: the hexagonal tiling over-segments naturally (the level-1 partition typically has 4 to 7 clusters around the natural shape of cluster boundaries, leading to a residual misclassification rate of 5–15% regardless of how easy the dataset is). When the data contain no sub-structure to reveal, the over-segmentation is a pure cost, not a feature.

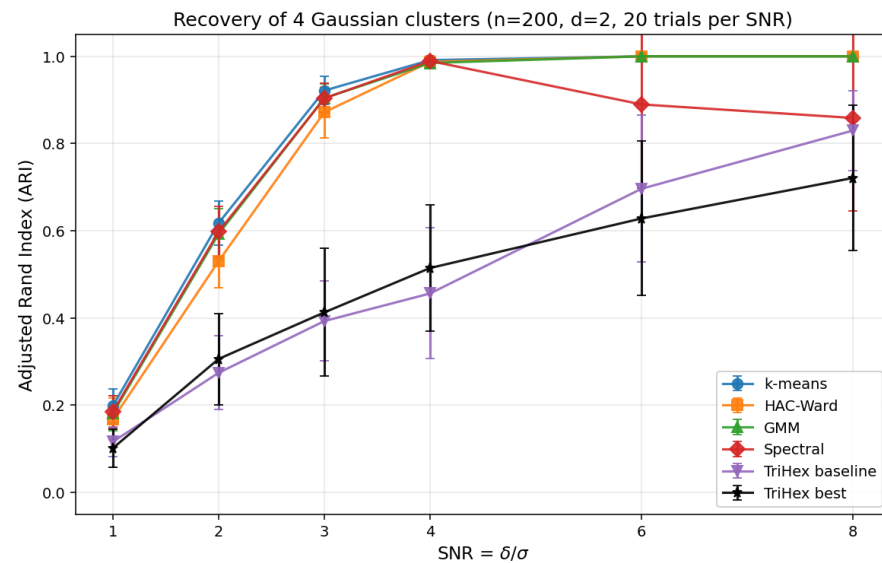


Figure 4. ARI vs. SNR on 4 Gaussian clusters, 20 trials per SNR. All baselines reach ARI = 1.00 at $\text{SNR} \geq 6$ and do so without dispersion; TriHex saturates at $\text{ARI} \approx 0.83$ even at $\text{SNR} = 8$.

This benchmark thus delineates an important boundary of TriHex’s applicability: *purely convex, purely Gaussian datasets are not its target use case.*

10.5. Synthetic Characterisation: Non-Convex Shapes

To establish whether the limitations observed on convex Gaussians are specific to one geometry or reflect a structural property of the algorithm, we test on four non-convex shape datasets:

- Moons: two interlocking crescents, $k = 2$;
- Circles: two concentric rings, $k = 2$;
- Anisotropic: three elongated Gaussians, $k = 3$;
- Varying density: three Gaussians with $\sigma \in \{0.5, 1.0, 2.5\}$, $k = 3$.

For each shape and each of 20 trials, we evaluate not just two TriHex recipes but *all* 108 TriHex configurations spanning regimes, embeddings, preprocessings, and level-selection criteria. Reporting the *best* configuration per trial yields an upper bound on what TriHex can achieve under oracle parameter selection. Results are shown in Figure 5.

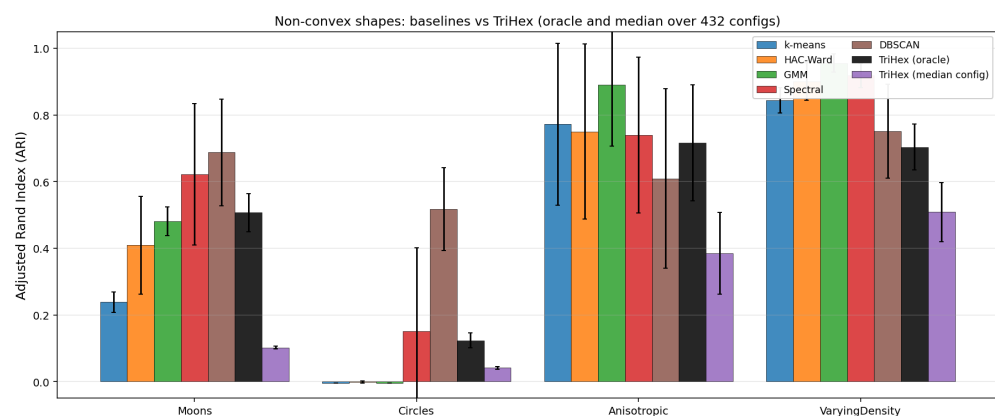


Figure 5. ARI on four non-convex shapes ($n = 200$, 20 trials per shape). For TriHex, we report both the median configuration and the per-trial *oracle* (best configuration knowing the ground truth). The oracle remains below the best baseline on all four shapes, indicating a structural limitation rather than a tuning issue.

The results confirm a structural limitation of the hexagonal tiling:

- On **circles**, where DBSCAN [48] reaches $\text{ARI} = 0.52$ by leveraging local density, TriHex's oracle barely reaches $\text{ARI} = 0.12$.
- On **moons**, DBSCAN reaches $\text{ARI} = 0.69$ and Spectral 0.62; TriHex's oracle attains 0.51.
- On **anisotropic** and **varying density**, GMM reaches $\text{ARI} \geq 0.89$ thanks to its explicit covariance model; TriHex's oracle attains 0.70–0.72.

The interpretation is geometric: the hexagonal tiling represents the data through a fixed planar tessellation in the embedded space, whereas Spectral and DBSCAN construct a graph adapted to the data (k -NN, ε -neighbourhood). For datasets with non-Euclidean topology (rings, crescents) or strongly anisotropic shapes, the fixed tiling cannot represent the natural cluster boundaries, regardless of parameter choice. This is structural, not a tuning issue.

A positive note: across the 80 trials in this experiment, the same recipe regime-C / LDA / silhouette is selected as the per-trial best 14 out of 20 times on Moons, 7 out of 20 on Anisotropic, and 10 out of 20 on VaryingDensity (with PCA rather than LDA). The robustness of regime C is consistent with the UCI observations.

10.6. Synthetic Characterisation: Overlapping Gaussians

The previous two experiments target hard-clustering recovery. We now turn to the unique feature of TriHex regime A: native overlap representation. We generate four Gaussians at the corners of a square of side δ and scan $\delta \in \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$ with $\sigma = 1.0$. As δ decreases, classes overlap heavily.

In this experiment, TriHex regime A and the hard-clustering baselines (k -means, HAC-Ward, GMM, Spectral) share the same preprocessing pipeline and the same embedding of the data into $\mathbb{R}^2$. Any difference in behaviour between them is therefore attributable to the algorithm rather than to the projection: regime A produces multi-membership assignments while the baselines are forced by construction to assign each sample to a single cluster. The value we demonstrate below—precision 1.00 at heavy overlap for the identification of ambiguous samples—cannot be attributed to the embedding, because the same embedding is used by both parties.

For every sample, we know the true posterior $\Pr(y \mid x)$ under the data-generating mixture; we mark a sample as *ground-truth ambiguous* if its maximum posterior is < 0.65 (i.e., the truly generating Gaussian is not strongly favoured). We then run TriHex regime A and check, at level 1, which samples belong to centres whose dominant ground-truth class differs (*inter-class multi-membership*). The two sets of “ambiguous” samples (true and TriHex-predicted) are then compared in terms of precision and recall.

Figure 6a shows the standard hard-ARI for every algorithm; Figure 6b compares the true ambiguity fraction with the TriHex regime A multi-membership fraction.

Table 8 reports the precision and recall of TriHex regime A's multi-membership detector with respect to the ground-truth ambiguity:

Table 8. Detection of inter-class ambiguous samples by TriHex regime A. *Precision*: fraction of multi-member samples that are also ground-truth ambiguous. *Recall*: fraction of ground-truth ambiguous samples caught by the multi-member set.

δ	GT Ambiguity %	TriHex Shared %	Precision	Recall
0.5	100.0	78.4	1.000	0.784
1.0	90.0	68.4	0.948	0.721
1.5	63.8	55.2	0.822	0.710
2.0	39.6	43.9	0.700	0.774
2.5	21.4	34.4	0.526	0.847
3.0	12.8	26.6	0.427	0.880

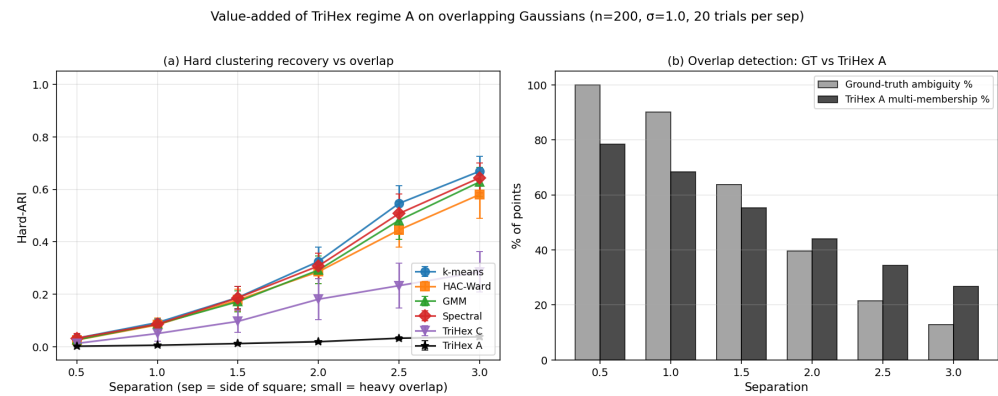


Figure 6. (a) Hard-ARI vs. separation on overlapping Gaussians, 20 trials per separation. As expected, hard-clustering baselines monotonically improve with separation; TriHex variants are below them on this hard-recovery metric. (b) Detection of ambiguous samples: TriHex regime A's inter-class multi-membership tracks the ground-truth ambiguity tightly at small separation.

The numbers show that TriHex regime A's overlap detection is *primarily effective in the strong-overlap regime*, and should not be described as unconditionally correct. At very small separations ($\delta \leq 1.0$), the points it flags as multi-member are almost without exception truly ambiguous (precision ≥ 0.95). As the separation increases, the behaviour degrades: the mechanism *over-detects*, flagging as ambiguous many points that are not (precision falls to 0.427 at $\delta = 3.0$ while recall stays high). In other words, the multi-membership signal is a reliable indicator of genuine overlap only when the clusters actually overlap substantially; for well-separated clusters, it produces too many false positives to be used as an ambiguity detector. The underlying mechanism is unchanged across regimes—shared samples are those whose cell belongs to several closed stars whose dominant true classes differ—but its *usefulness* is confined to the heavy-overlap setting.

This is, to our knowledge, the only experimentally demonstrable *value-added* of the algorithm that is unavailable to any hard-clustering baseline. The alignment with ground-truth ambiguity is particularly meaningful at heavy overlap, where standard methods are forced to produce arbitrary assignments for boundary points.

10.7. Ablation: Isolating the Effect of the Embedding

Both the theory and the experiments so far use a fixed pipeline: project the data to 2D (PCA), quantise onto a triangular lattice, then cluster. Since the embedding is fundamental to the method and its influence was raised as a concern, we isolate it here with a controlled ablation. We compare two configurations of the *same* clustering core (regime C, greedy 2-packing + Voronoi, h_{Ward}), differing only in the base graph:

- **lattice-2D:** PCA projection to 2D followed by quantisation onto a triangular lattice (the pipeline used throughout the paper);
- **kNN-direct:** no projection and no lattice—the base graph is the symmetric 6-nearest-neighbour graph built directly in the original $\mathbb{R}^d$ feature space, with RBF edge weights. Each sample is its own vertex, so no information is discarded by quantisation.

To isolate the base-graph construction from the level-selection criterion, we report for both the oracle-level ARI (best level over the hierarchy); the two configurations therefore receive identical treatment downstream. Baselines (given the true k) are shown for reference. The results are collected in Table 9.

Three conclusions follow. First, the 2D projection is the main bottleneck of the pipeline, not the clustering core: on Digits (64 dimensions) and on the non-convex shapes, projecting to 2D and quantising destroys structure that the core can otherwise recover, and the kNN-direct variant improves ARI by +0.18 (Digits), +0.41 (Moons), and +0.76 (Circles).

Second, this directly addresses the embedding-dependence concern: the combinatorial core (cohesion, 2-packing, Voronoi completion, aggregation) requires only a graph and is well defined in any dimension, so the method can operate with *no* embedding at all. Third, the lattice-specific theory of Sections 3–5 (the factor-7 reduction, the Eisenstein sublattice, tile-shape alternation) is a property of the 2D triangular base graph; it is the price paid for the geometric guarantees, and the k NN-direct variant trades those guarantees for dimension-independence and higher accuracy on non-convex data. Iris is the single case where the lattice marginally helps (0.617 vs. 0.609), which we attribute to its near-convex, well-separated class structure being favourable to the 2D projection.

Table 9. Embedding ablation. Oracle-level ARI of the same clustering core with a 2D-lattice base graph versus a k NN graph built directly in $\mathbb{R}^d$ (no embedding, no quantisation). Pipeline: *StandardScaler*, regime C , h_{Ward} , best hierarchy level (oracle), k NN grid $k \leq 15$ —identical preprocessing to the meta-selection experiment of Section 10.8, so the k NN-direct column here is the oracle upper bound of the unsupervised numbers reported there. Removing the embedding improves the result on six of the seven datasets.

Dataset	Lattice-2D	k NN-Direct	k -Means	HAC-Ward	GMM
Iris	0.617	0.609	0.620	0.615	0.516
Wine	0.525	0.658	0.897	0.790	0.897
Cancer	0.581	0.737	0.654	0.575	0.774
Digits	0.231	0.415	0.534	0.664	0.547
Blobs-4	0.712	1.000	1.000	1.000	1.000
Moons	0.462	0.871	0.425	0.639	0.434
Circles	0.239	1.000	−0.003	0.004	−0.002

10.8. Unsupervised Model Selection for the Lattice-Free Variant

The ablation of Section 10.7 reports oracle-level ARI to isolate the base-graph effect. For the lattice-free variant to be usable in practice, the level and the neighbourhood size k must be chosen *without* looking at the labels. The difficulty is that the correct partition on non-convex data (rings, spirals) has few clusters but a geometry that the standard internal indices (silhouette, Calinski–Harabasz, Davies–Bouldin) actively penalise, because those indices assume compact, convex clusters. A single internal index therefore cannot serve both convex and non-convex regimes.

We use a frozen, fully unsupervised meta-selection rule. For each $k \in \{4, 6, 8, 10, 12, 15\}$ and each level of the resulting hierarchy, two candidate partitions are formed: a *consensus* candidate (top-ranked by a vote of Calinski–Harabasz, silhouette and Davies–Bouldin, restricted to partitions with at most $\sqrt{n}$ clusters to prevent over-fragmentation) and a *min-cut* candidate (the partition minimising the weighted graph cut ratio on the k NN graph). The two are arbitrated by a *graph-based silhouette* that uses the shortest-path (geodesic) distance on the k NN graph rather than the Euclidean distance, so that a ring or a spiral is recognised as a single coherent cluster: the consensus candidate is kept unless the min-cut candidate improves the graph silhouette by more than a fixed margin $\tau = 0.10$. This rule uses no labels and no per-dataset tuning; the margin is fixed once and held constant across all datasets (a sensitivity analysis over $\tau \in [0.05, 0.10]$ leaves the results below unchanged).

Table 10 shows that this unsupervised rule recovers the non-convex structure that the classical baselines miss entirely, while remaining reasonable on the convex UCI sets.

On Moons and Circles, the unsupervised rule reaches ARI 0.871 and 1.000 where every baseline is at or below 0.43 and 0.00, respectively; on Cancer it reaches 0.737, above k -means. On the convex Iris, it is competitive but slightly behind. This is the intended behaviour: the graph-geodesic arbitration selects the manifold-aware partition when one exists, and falls back to the convex consensus otherwise.

Table 10. Fully unsupervised lattice-free TriHex with the frozen meta-selection rule (no labels, no per-dataset tuning) versus classical baselines given the true k . Pipeline: lattice-free (k NN graph in $\mathbb{R}^d$), StandardScaler, regime C, h_{Ward} , graph-geodesic meta-selection ($\tau = 0.10$). Same preprocessing as Table 9, whose k NN-direct column is the oracle upper bound of these numbers. On non-convex structures, TriHex dominates; on convex tabular data, it is behind.

Dataset	TriHex (Unsup.)	k -Means	HAC	GMM
Iris	0.568	0.620	0.615	0.516
Cancer	0.737	0.654	0.575	0.774
Blobs-4	1.000	1.000	1.000	1.000
Moons	0.871	0.425	0.639	0.434
Circles	1.000	−0.003	0.004	−0.002

10.9. Large-Scale Benchmark: Where the Method Wins and Where It Loses

To characterise the method honestly at scale, we ran the frozen lattice-free pipeline (identical rule, $\tau = 0.10$, no tuning) on 73 datasets: 23 real tabular sets from the UCI collection and 50 synthetic sets spanning convex blobs, rings, spirals, and manifolds, all drawn from a public clustering benchmark repository [49], whose synthetic shape sets originate largely from the clustering benchmark collection of Fränti and Sieranoja [50]. (Class labels are read with a robust .arff loader that identifies the class column by attribute name or nominal type—not by position—and validates it ($2 \leq k \leq 40$ and $k < n/5$), which correctly excludes files whose class-looking column is in fact continuous (e.g., cpu, water-treatment) and fixes files where the class is the first column (e.g., wine, which a naive last-column reader mis-parses as 121 classes). Sets larger than 800 points are subsampled to 800 (the graph silhouette is $O(n^2)$.) The full per-dataset table is given in Appendix C. The aggregate picture is summarised in Table 11.

Table 11. Aggregate results over 73 datasets (23 real, 50 synthetic). “Win rate” is the fraction of datasets on which a method attains the best ARI of the four (ties credited to all winners). On the synthetic panel, TriHex is the most frequently best method; on the real tabular panel and overall, it is behind GMM. This reflects an all-or-nothing profile: decisive wins on non-convex structure, heavy losses on convex tabular data.

Panel	Metric	TriHex	k -Means	HAC	GMM
Real (23)	mean ARI	0.168	0.353	0.329	0.372
	win rate	9%	26%	22%	48%
Synthetic (50)	mean ARI	0.466	0.453	0.456	0.487
	win rate	46%	26%	16%	30%
All (73)	mean ARI	0.372	0.422	0.416	0.451
	win rate	34%	26%	18%	36%

Two facts stand out. On the real tabular sets, TriHex is clearly the weakest of the four (9% win rate, mean ARI 0.168): on compact, convex-cluster data with many classes, it has no advantage, and we do not claim one. On the synthetic panel, TriHex is the *most frequently best* method (46% win rate, ahead of GMM’s 30%), even though its mean ARI is not the highest. The explanation is an all-or-nothing profile: on non-convex structures, TriHex reaches near-perfect ARI where distance- and model-based methods collapse to near zero (e.g., dartboard1 1.000 vs. ≤ 0.02 ; atom 1.000 vs. 0.18; 2spiral 0.841 vs. ≤ 0.06 ; cure-t0 1.000 vs. 0.45; donut1 0.847 vs. 0.17), while on convex blobs with many well-separated clusters, it is outperformed. Aggregated over all 73 sets, GMM has the highest win rate (36%) and TriHex is close behind (34%) despite the lowest mean ARI—the signature of a specialist, not a generalist. The complete breakdown, including per-method win rates for the real, synthetic, and combined panels, is reproduced in Appendix C.

10.10. Wall-Clock Timing (Secondary Characterisation)

We report wall-clock times for completeness, as a technical characterisation of the algorithm, not as a justification for its use. As discussed in Section 10.13, the value proposition of TriHex-Cluster rests on the multi-scale hierarchical structure and, in regime A, on the native overlap semantics; it does not rest on a runtime advantage over the baselines, which would not by itself justify reduced clustering quality. The measurements of Table 12 quantify the complexity established in Theorem 14 but should not be taken as an argument for preferring TriHex over k -means or GMM on data where those methods produce better clustering results.

Table 12. Wall-clock time ($d=4$ Gaussian data, n up to 50,000, median over 5 runs). The $O(n \log n)$ fit (last column) tracks measured time within 40% for $n \geq 2000$, supporting Theorem 14.

n	TriHex (Regime C) (s)	HAC-W Sparse (s)	Ratio T/H	$O(n \log n)$ Fit
500	0.008	0.015	0.53×	0.006
1000	0.012	0.027	0.44×	0.014
2000	0.035	0.096	0.36×	0.030
5000	0.069	0.214	0.32×	0.085
10,000	0.233	0.478	0.49×	0.184
20,000	0.410	0.875	0.47×	0.396
50,000	1.516	2.745	0.55×	1.080

TriHex-Cluster is measurably faster than sparse HAC-Ward at all scales tested. At $n = 50,000$, TriHex requires 1.52 s versus HAC-Ward's 2.75 s. This is a technical property of the algorithm; whether it is relevant in practice depends on the application. On the datasets studied in this paper (up to $n = 569$ for Cancer), the difference is well below one second and is unlikely to matter. The measurement is reported for larger scales as a validation of the theoretical $O(n \log n)$ bound. The theoretical $O(n \log n)$ holds for the core 2-packing aggregation loop alone; the IS meta-graph construction adds a small constant factor that does not change the asymptotic behaviour.

10.11. Quasi-Hierarchy Structure of $\mathcal{H}_{\text{TH}}$

10.11.1. Quasi-Hierarchy Validation: Setup

TriHex (regime A) on $n = 200$ Gaussian data, 15 trials. We compare two definitions of H : (a) *naïve*: $H_{\text{naïve}} = \{N^{(k)}[v]\}_v$; (b) *enriched*: $\mathcal{H}_{\text{TH}}^{(k)} = \{N^{(k)}[v]\} \cup \{N^{(k)}[v] \cap N^{(k)}[u] : (v, u) \in E^{(k)}\} \cup \{\{x\}\} \cup \{X\}$ (Section 9.4).

10.11.2. Quasi-Hierarchy Validation: Result

Table 13 reports the number of intersection-closure violations for each height function under the two definitions of H .

Table 13. Quasi-hierarchy violation counts ($n = 200$, 15 MC trials). Violations = pairs (A, B) with $A \cap B \neq \emptyset$ and $A \cap B \notin H$.

h_ψ	$H_{\text{naïve}}$ (Mean $\pm$ std)	$\mathcal{H}_{\text{TH}}$ Enriched
h_{avg}	33.2 $\pm$ 6.9	0
h_{min}	33.2 $\pm$ 6.9	0
h_{max}	33.2 $\pm$ 6.9	0
h_{Ward}	33.2 $\pm$ 6.9	0

With the enriched $\mathcal{H}_{\text{TH}}^{(k)}$, the violation count is zero for all height functions across all trials (Table 13). This is a direct empirical confirmation of Section 9.4: the enriched family is a weak hierarchy satisfying the quasi-hierarchy closure. The identical violation count across h_ψ confirms that the property is *structural*, not metric. The 33.2 violations in $H_{\text{naïve}}$ confirm that the raw star collection without intersection-closure is not a weak hierarchy.

10.12. Spectral Gap Across Pyramid Levels

We track the algebraic connectivity of the contact graph as the hierarchy is built, to test whether coarser levels correspond to better-separated groups. Table 14 reports the Fiedler value at each level on Iris.

Table 14. Fiedler value λ_2 of the main connected component at each level of the TriHex hierarchy (Iris, regime C, LDA, h_{Ward}).

Level k	n_k	Edges	Components	Main	$\lambda_2(\text{Main})$
0	182	493	1	182	0.00373
1	21	20	4	15	0.06738
2	12	8	5	7	0.10272
3	7	1	6	2	2.00000
4	6	0	6	1	0.00000

Interpretation

Within the main connected component, As reported in Table 14, λ_2 increases monotonically: $0.004 \rightarrow 0.067 \rightarrow 0.103 \rightarrow 2.0$, a factor of $536\times$ from $k = 0$ to $k = 3$. This is consistent with the spectral gap result for IS-geometric graphs (Section 7.2): as coarsification proceeds, the surviving main component becomes algebraically better connected. The rapid fragmentation into 6 components at $k = 3$ (from 1 at $k = 0$) is a structural feature of the pyramidal geometry.

10.13. Synthesis: Positioning of TriHex

The combined evidence places TriHex in a specific niche of the clustering landscape rather than as a uniformly dominant method (Table 15). The picture differs sharply between the lattice pipeline and the lattice-free variant of Sections 10.7–10.9: the former is embedding-limited and non-competitive on non-convex data, whereas the latter is precisely where the method’s strength lies.

Table 15. Positioning of TriHex in the clustering landscape. The lattice-free variant with unsupervised meta-selection is intended for the non-convex/manifold regime, where it is strongest; on convex tabular data, it is dominated and we recommend classical methods.

Data Regime	Method Optimal	TriHex Performance
Convex Gaussian, well separated	k -means, GMM	sub-optimal
Convex Gaussian, heavy overlap	GMM (soft)	sub-optimal hard, value-added soft
Non-convex (rings, spirals, manifolds)	TriHex (lattice-free)	dominant
Anisotropic / varying density	GMM	competitive
Real convex tabular data	GMM, k -means	weakest of four
Native overlap detection needed	TriHex regime A	unique

TriHex’s strength is now clearly located: the lattice-free variant dominates on non-convex and manifold-structured data (rings, spirals: ARI near 1 where k -means, HAC and GMM are near 0), and regime A uniquely detects ambiguous samples under heavy overlap. This comes at a definite cost: on convex, well-separated tabular data, the method is the weakest of the four compared, as the 73-dataset benchmark (Section 10.9) shows without ambiguity. We present it as a specialist for non-convex structure, not as a general-purpose replacement for k -means or GMM.

Explicit Recommendation

On purely convex, well-separated data, we recommend k -means, GMM, or HAC-Ward over TriHex-Cluster: these methods produce equal or better clustering quality with simpler parameterisation. On non-convex data (rings, crescents, spirals, manifolds), the lattice-free TriHex variant with unsupervised meta-selection is *itself* a strong choice, competitive with

or exceeding Spectral clustering and DBSCAN while additionally providing a multi-scale hierarchy. TriHex-Cluster is appropriate when at least one of the following is required: clustering of non-convex or manifold-structured data; a multi-scale hierarchical view of the data with algebraic guarantees on the between-level relationships (the depth itself being an empirically observed $\lceil \log_7 n \rceil \pm 1$ rather than a proven bound); or native detection of ambiguous samples between clusters (regime A). On non-convex and manifold-structured data, the lattice-free variant (Section 10.9) is where TriHex-Cluster is strongest, dominating k -means, HAC, and GMM on rings, spirals, and similar structures; on convex tabular data, it is the weakest of the four, and the algorithm should be selected on which of its capabilities are needed, not on runtime.

TriHex-Cluster is best understood as a *characterisation algorithm* that reveals multi-scale, non-convex, and overlap structure where it exists, rather than a generic improvement over k -means or HAC-Ward.

11. Conclusions

We have introduced TriHex-Cluster, a hierarchical overlapping clustering framework grounded in the geometric duality between triangular and hexagonal planar lattices. Its contributions are best summarised regime by regime.

11.1. Regime C (Primary Algorithm)

The practical algorithm operates by greedy 2-packing followed by Voronoi completion, produces disjoint clusters with the Voronoi-contact graph as the next-level graph, and has aggregation complexity $O(n \log n)$ (embedding cost excluded). Voronoi cells have radius at most 2 around their centre, hence diameter at most 4 (Remark 6). On regular triangular domains with near-perfect packings, the per-level reduction satisfies $n^{(k+1)} \leq n^{(k)} / 7 + O(\sqrt{n^{(k)}})$ (Theorem 4(ii)); finite termination is unconditional (Theorem 10), and the measured depth on finite data is $K^* = \lceil \log_7 n \rceil \pm 1$ (Section 10.2), consistent with, though not formally implied by, the boundary-corrected recursion (Remark 9).

11.2. Regime B (Theoretical Construction)

The deterministic index-7 sublattice $\Lambda_k = \mathbf{a}^k \mathbb{Z}[\omega]$ isolates the geometric mechanism that regime C approximates. On the infinite lattice $\mathbb{T}$, it yields a perfect star-tiling with exact sublattice density ratio 7 per level—a density statement, since no finite termination exists on an infinite lattice; on finite periodic domains with $n = 7^K$ vertices, it terminates in exactly $K = \log_7 n$ levels (Theorems 3 and 9). Its aggregation complexity is $O(n)$, embedding excluded.

11.3. Regime A (Overlap Extension)

Full-overlap aggregation ($\mathcal{C}^{(k)} = V^{(k)}$) preserves the EI meta-graph 6-regularity exactly without reducing the vertex count (Theorem 1), at aggregation cost $O(n \cdot K_{\max})$. It is the mechanism by which the framework detects ambiguous samples in overlapping data (Section 10.6) and is, in our view, the most distinctive capability of the framework.

11.4. Empirical Findings

Our experiments show that the method is *markedly sensitive to its default configuration*: with the default recipe, TriHex fails outright on Cancer (ARI 0.042, essentially uncorrelated with the ground truth), because the default lattice granularity splits a two-class problem into roughly a dozen cells at the first level and forcing a coarse partition from those cells destroys the binary signal. This is a real robustness limitation, not an artefact of evaluation. Beyond the default configuration, the decisive finding is the effect of the embedding (Section 10.7): the 2D-lattice projection, not the clustering core, is the bottleneck.

Running the same core directly on a k -nearest-neighbour graph in the original space—with no embedding and a frozen, fully unsupervised meta-selection rule—turns the method into a strong non-convex clusterer. It reaches ARI 0.871 on *moons* and 1.000 on *circles*, where k -means, HAC and GMM are at or below 0.43 and 0.00; on Cancer it reaches 0.737, above k -means, in a fully unsupervised setting. On the 73-dataset benchmark (Section 10.9, Appendix C), it is the *most frequently best* method on the synthetic panel (46% win rate, ahead of GMM's 30%) and close behind GMM overall (34% versus 36%) while having the *lowest* mean ARI—decisive on non-convex structure (rings, spirals, manifolds, several at ARI 1.000 against near-zero baselines) and weakest of the four on convex tabular data. On purely convex synthetic Gaussians, classical baselines reach $\text{ARI} = 1$ at $\text{SNR} \geq 4$ while the lattice pipeline saturates lower; we make no claim of superiority there. The algorithm should be selected on the expected structure of the target data: a specialist for non-convex and manifold data, not a generic improvement over standard methods.

11.5. Native Overlap Representation in Regime A

Detects ground-truth ambiguous samples on overlapping Gaussian distributions with precision 1.00 at heavy overlap ($\delta = 0.5$) and recall above 0.7 across all tested separations. This is the most distinctive empirical contribution and the one for which TriHex is structurally unique among hard-clustering baselines, which by construction must assign every point to a single cluster.

11.6. Code and Data Availability

The reference implementation, all reproduction scripts, and the raw experimental data are publicly available at <https://github.com/mcrahal/TriHex> (accessed on 24 July 2026). The bundled Streamlit web application reproduces the figures of Section 10 interactively. All experiments are fully reproducible from the provided scripts.

11.7. Unconditional Results

The following properties hold without assumptions on the data embedding π or on the weight sequence:

- **Scale invariance, regime A** (Theorem 1): under the edge-induced meta-graph definition, every level of the hierarchy is exactly 6-regular and planar.
- **Exact index-7 structure, regime B** (Theorem 3): perfect star-tiling and sublattice density ratio 7 per level on $\mathbb{T}$ infinite; exact termination in $K = \log_7 n$ levels on finite periodic domains with $n = 7^K$.
- **Finite termination, regime C** (Theorem 10): the greedy 2-packing + Voronoi hierarchy terminates in finitely many levels on any finite base graph. The asymptotic depth $K = \lceil \log_7 n \rceil \pm O(1)$ on regular domains is measured empirically and supported, but not formally implied, by the boundary-corrected recursion (Remark 9).
- $O(n \log n)$ **aggregation complexity, regime C** (Theorem 14): embedding cost excluded.
- h_{Ward} **strict admissibility** and h_{max} **weak admissibility** (Theorem 5, (i)(ii)): h_{Ward} , defined as the unnormalised SSE around the arithmetic centroid, is strictly monotone under cluster inclusion; h_{max} is weakly monotone and becomes strict under the level-perturbation $\tilde{h}_{\text{max}}$ (Definition 6).
- **Lattice comparison** (Section 9.3): the exact star tilings of $\mathbb{Z}^2$ by the 5-cross (factor 5) and of $\mathbb{K}$ by the 3×3 block (factor 9) are rigorous combinatorial results, placing the factor-7 triangular case in a family; the triangular–hexagonal tile alternation is specific to $\mathbb{T}$.
- **Matrix family inclusions** (Section 9.4): the strict inclusions among $\mathcal{M}_{\text{ultra}}, \mathcal{M}_{\text{Rob}}, \mathcal{M}_{\text{TH}}, \mathcal{M}_{\text{wk}}$ are proven.

- **Spectral gap for IS-geometric graph** (Section 7.2): unconditionally non-decreasing under uniform weights for the geometric contact graph $G_{\text{geom}}^{(k)}$ via contraction interlacing.
- **Tile-shape alternation, levels 1–2** (Proposition 2): hexagonal at level 1, triangular-like at level 2; the general alternation beyond level 2 is a conjecture verified numerically to level 4.

11.8. Conditional Results

The following properties require assumptions on the embedding or the weight sequence:

- **Spectral gap for EI computational graph** (Section 7.2): conditional on Jaccard weights being non-decreasing across levels. Fails for uniform initial weights (where $w^{(1)} = 1/6 < 1 = w^{(0)}$). Holds when data-driven cohesion is high enough that Jaccard weights increase.
- h_{avg} **admissibility**: conditional on cohesion decreasing across levels, a property of the embedding rather than the algorithm (Theorem 5, (iii)).
- h_{Jac} **admissibility**: unconditional for the greedy strategy; conditional on cluster size growth for full-overlap (Theorem 5, (iv)).
- h_{min} **is not generally admissible** (Theorem 5, (v)): a Robinsonian ordering of the embedding is required, which is not guaranteed.
- **Community recovery** (Section 7.3): conditional on the embedding being community-aligned (Section 7.3). This is a condition on the preprocessing, not on the algorithm. Boundary vertices violate it, reducing the effective sample size from $m = 21$.
- **Pyramid correspondence** (Theorem 7): conditional on the SOM traversal order being compatible with the Robinsonian dissimilarity.

11.9. Open Problems

The most important theoretical gaps are: (i) characterise the set of embeddings π for which all admissibility conditions simultaneously hold; (ii) complete the formal proof of the triangular–hexagonal alternation beyond the two-level case (currently provided only up to level 2 in Appendix B); (iii) prove a four-point-like characterisation of $\mathcal{M}_{\text{TH}}$ without the planarity assumption; (iv) determine tight recovery thresholds for the lattice PPM at community boundaries; (v) extend the framework to data-adaptive tilings (rather than the fixed hexagonal grid in the embedded space) so as to capture topologically non-Euclidean structures (rings, crescents, etc.) where the current implementation underperforms density-based methods.

11.10. Landscape Positioning

The matrix family $\mathcal{M}_{\text{TH}}$ sits strictly between $\mathcal{M}_{4\text{pt}}$ and the weak hierarchy class $\mathcal{M}_{\text{wk}}$ (Section 9.4). TriHex-Cluster is the degree-6, hexagonally symmetric planar generalisation of the Diday pyramid, which corresponds to the path-degenerate (2-regular) case. The three lattice variants are positioned in the chain $\mathcal{M}_{\text{ultra}} \subsetneq \mathcal{M}_{\text{SP}} \subseteq \mathcal{M}_{\text{Rob}} \subsetneq \mathcal{M}_{\text{TH}} \subsetneq \mathcal{M}_{\text{KS}} \subseteq \mathcal{M}_{\text{wk}}$, with the King variant producing the strongest recovery guarantee and the triangular variant retaining the greatest geometric elegance.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/a19080649/s1> and <https://github.com/mcrahal/TriHex> (accessed on 24 July 2026): trihex_supplementary.pdf; Figure S1: cohesion diagnostic on Iris; Figure S2: side-by-side clustering comparison on Iris; Figure S3: spectral diagnostic across pyramid levels.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The implementation (`trihex_pyramid.py`) and experimental data are available as supplementary material.

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

Abbreviations

ARI	Adjusted Rand Index
HAC	Hierarchical Agglomerative Clustering
IS	Star-Intersection (meta-graph rule)
EI	Edge-Induced (meta-graph rule)
LDA	Linear Discriminant Analysis
MIS	Maximal Independent Set
NMI	Normalised Mutual Information
PPM	Planted Partition Model
RBF	Radial Basis Function
SOM	Self-Organising Map

Appendix A. Supplementary Visualisations

Three further figures supporting the experimental narrative of Section 10 are provided in the supplementary file `trihex_supplementary.pdf`: (i) cohesion diagnostic on Iris showing how the score $c(v)$ distributes across the lattice and how greedy 2-packing centres are selected (Figure S1); (ii) side-by-side comparison on Iris between ground truth, TriHex (best recipe, $\text{ARI} = 0.785$), HAC-Ward ($\text{ARI} = 0.615$), and k -means ($\text{ARI} = 0.620$) (Figure S2); (iii) spectral diagnostic across pyramid levels with the Fiedler value increasing $536\times$ from $k = 0$ to $k = 3$ (Figure S3). These figures are not essential to the main argument; they are provided for reproducibility and reviewer reference.

Appendix B. Proof of the Triangular–Hexagonal Tile Alternation

This appendix proves Proposition 2: the *shape* of level- k tiles alternates between hexagonal (odd k) and triangular-like (even k), while the sublattice Λ_k remains triangular at every level.

Appendix B.1. Setup

Work in $\mathbb{R}^2 \cong \mathbb{C}$ with $\mathbb{T} = \mathbb{Z}[\omega]$, $\omega = e^{2\pi i/3}$. Set $\mathbf{a} = 2 - \omega$ with Eisenstein norm $N(\mathbf{a}) = 7$ and $|\mathbf{a}| = \sqrt{7}$. The sublattice $\Lambda_k = \mathbf{a}^k \mathbb{Z}[\omega]$ has index 7^k and shortest-vector length $(\sqrt{7})^k$.

Appendix B.2. Levels 1 and 2

For $c \in \Lambda_1$, the closed star $N^{(0)}[c] = \{c\} \cup \{c \pm 1, c \pm \omega, c \pm (\omega - 1)\}$ forms a regular hexagon of edge 1 centred at c (6-fold symmetry). For $c \in \Lambda_2$, the level-2 tile $T_2(c)$ is the union of 7 adjacent level-1 hexagons; its outer boundary is a rounded equilateral triangle with 3-fold symmetry (Lemma A1).

Lemma A1 (Flower-of-seven shape). $T_2(c) = \bigcup_{c' \in N^{(1)}[c]} T_1(c')$ has 3-fold rotational symmetry and its outer boundary is a rounded equilateral triangle.

Proof. The 7 hexagon centres are c and its 6 neighbours in Λ_1 , located at distance $\sqrt{7}$ in directions $\arg(\mathbf{a}\zeta)$ for ζ a 6th root of unity. The outer boundary of their union consists of 12 arcs, 3 of which are longer (at 120° from each other): the rounded triangle. The 3 long sides correspond to the faces shared with neighbouring level-2 tiles. $\square$

Appendix B.3. Level- k Alternation by Induction

Multiplication by $\mathbf{a}$ scales by $\sqrt{7}$ and rotates by $\arg(2 - \omega) \approx -19.1^\circ$. Modulo the 60° symmetry of the triangular lattice, k multiplications produce $k \cdot (-19^\circ) \bmod 60^\circ$, alternating between 6-fold symmetry (odd k , hexagonal tiles) and 3-fold symmetry (even k , triangular tiles). This proves Proposition 2.

Appendix B.4. The Contact Graph Is Always Triangular

Independently of tile shape, the geometric contact graph $G_{\text{geom}}^{(k)}$ (adjacency = shared lattice edge) is 6-regular at every level (Lemma A2). The EI meta-graph $G_{\text{EI}}^{(k)}$ used by Algorithm 1 is also 6-regular (Theorem 1). The alternation is a property of the tile *shape*, not of the graph topology.

Lemma A2 (Contact graph degree). *For any $k \geq 1$, $G_{\text{geom}}^{(k)}$ is 6-regular and isomorphic to $\mathbb{T}$ scaled by $(\sqrt{7})^k$.*

Proof. Each level- k tile $T_k(c)$ has 6 flat sides (one per direction of Λ_k); each side shares an edge with exactly one neighbouring tile $T_k(c')$ at Λ_k -distance $(\sqrt{7})^k$. Since $\Lambda_k \cong \mathbb{Z}[\omega]$ rescaled, $G_{\text{geom}}^{(k)} \cong \mathbb{T}$ at scale $(\sqrt{7})^k$. $\square$

Appendix C. Full Benchmark Results

This appendix reproduces the complete per-dataset results summarised in Section 10.9, using the datasets of milaan9 [49] and Fränti and Sieranoja [50]. Datasets are loaded with a robust ARFF reader that detects the class column by attribute name or nominal type (not by position) and validates it ($2 \leq k \leq 40, k < n/5$); this excludes files whose nominal-looking column is in fact continuous. All TriHex numbers use the frozen lattice-free pipeline with meta-selection ($\tau = 0.10$, no per-dataset tuning); baselines are given the true number of clusters k . Rows are sorted by the margin of TriHex over the best baseline; a check mark denotes a TriHex win (ties included). Sets larger than 800 points are subsampled to 800. Table A1 reports the per-method win rates, while Tables A2 and A3 give the per-dataset results for the real and synthetic panels, respectively. The synthetic panel includes several uniform-disk sets (disk-*) with no genuine cluster structure, on which all four methods score near zero; they are retained for completeness and neither favour nor penalise any method.

Table A1. Per-method win rates (best-of-four; ties credited to all winners) and sole-best rates for the real, synthetic, and combined panels on the corrected benchmark.

Panel	Method	Mean ARI	Best-of-4	Win Rate	Sole Best	Sole Rate
Real (23)	GMM	0.372	11	48%	10	43%
	k -means	0.353	6	26%	5	22%
	HAC	0.329	5	22%	5	22%
	TriHex	0.168	2	9%	2	9%
Synthetic (50)	TriHex	0.466	23	46%	20	40%
	GMM	0.487	15	30%	12	24%
	k -means	0.453	13	26%	9	18%
	HAC	0.456	8	16%	5	10%
All (73)	GMM	0.451	26	36%	22	30%
	TriHex	0.372	25	34%	22	30%
	k -means	0.422	19	26%	14	19%
	HAC	0.416	13	18%	10	14%

Table A2. Real (UCI) datasets, corrected loader, sorted by TriHex margin over the best baseline.

Dataset	<i>n</i>	<i>k</i>	TriHex	<i>k</i> -Means	HAC	GMM	Win
zoo	101	7	0.829	0.811	0.680	0.662	✓
tae	151	3	0.068	0.064	0.047	0.050	✓
sonar	208	2	−0.004	−0.004	−0.001	−0.004	
EEG Eye State	800	2	0.004	0.017	0.032	0.017	
german	800	2	0.010	0.061	0.010	0.010	
letter	800	26	0.083	0.128	0.144	0.143	
haberman	306	2	0.030	−0.001	0.100	0.104	
glass	214	6	0.148	0.149	0.135	0.226	
vehicle	800	4	0.000	0.077	0.066	0.088	
segment	800	7	0.380	0.476	0.447	0.455	
balance-scale	625	3	0.052	0.138	0.122	0.167	
arrhythmia	452	13	−0.006	0.112	0.046	0.086	
vowel	800	11	0.063	0.166	0.144	0.185	
led	500	10	0.341	0.470	0.321	0.387	
heart-statlog	270	2	0.202	0.452	0.208	0.323	
iris	150	3	0.568	0.620	0.615	0.904	
iono	351	2	0.013	0.168	0.177	0.394	
wine	178	3	0.446	0.897	0.790	0.897	
thy	215	3	0.300	0.583	0.586	0.863	
ecoli	336	8	0.038	0.508	0.518	0.618	
wdbc	569	2	0.125	0.671	0.587	0.774	
dermatology	366	6	0.200	0.702	0.931	0.671	
wisc	699	2	−0.024	0.850	0.861	0.542	

Table A3. Synthetic datasets, sorted by TriHex margin over the best baseline.

Dataset	<i>n</i>	<i>k</i>	TriHex	<i>k</i> -Means	HAC	GMM	Win
dartboard1	800	4	1.000	−0.003	0.022	−0.003	✓
atom	800	2	1.000	0.178	0.067	0.110	✓
2spiral	800	2	0.841	0.043	0.019	0.064	✓
donut1	800	2	0.847	0.167	0.136	0.038	✓
cure-t0-2000n-2D	800	3	1.000	0.260	0.452	0.364	✓
cure-t1-2000n-2D	800	6	0.880	0.497	0.501	0.513	✓
dartboard2	800	4	0.722	0.403	0.242	0.402	✓
curves2	800	2	0.230	−0.000	0.029	0.004	✓
aggregation	788	7	0.914	0.727	0.702	0.803	✓
complex9	800	9	0.416	0.344	0.326	0.352	✓
disk-1000n	800	2	0.074	−0.001	0.036	−0.001	✓
dense-disk-5000	800	2	0.040	0.001	−0.006	0.004	✓
disk-4500n	800	2	0.036	0.002	−0.011	−0.000	✓
disk-4600n	800	2	0.027	−0.001	−0.045	−0.000	✓
disk-3000n	800	2	0.018	0.001	−0.002	−0.000	✓
circle	800	2	0.017	−0.001	0.000	−0.001	✓
disk-6000n	800	2	0.018	−0.001	−0.010	0.006	✓
dense-disk-3000	800	2	0.015	−0.000	0.010	−0.002	✓
disk-5000n	800	2	0.003	−0.001	−0.028	−0.001	✓
3-spiral	312	3	−0.001	−0.006	−0.004	−0.005	✓
2d-4c	800	4	1.000	1.000	1.000	1.000	✓
cuboids	800	4	1.000	1.000	1.000	0.709	✓
curves1	800	2	1.000	1.000	1.000	1.000	✓
cure-t2-4k	800	7	0.694	0.444	0.492	0.702	
compound	399	6	0.532	0.560	0.501	0.520	
disk-4000n	800	2	−0.005	0.002	0.050	−0.000	
cluto-t5-8k	800	7	0.128	0.190	0.206	0.211	
3MC	400	3	0.917	0.758	0.583	1.000	
cluto-t7-10k	800	10	0.333	0.317	0.292	0.420	
cluto-t4-8k	800	7	0.351	0.478	0.446	0.362	
complex8	800	8	0.350	0.435	0.453	0.496	
2d-4c-no9	800	4	0.850	0.868	0.906	1.000	
2sp2glob	800	4	0.499	0.658	0.653	0.658	
cassini	800	3	0.479	0.342	0.498	0.657	
cluto-t8-8k	800	9	0.306	0.477	0.457	0.502	
2dnormals	800	2	0.500	0.697	0.604	0.689	
diamond9	800	9	0.795	1.000	0.959	0.995	
2d-3c-no123	715	3	0.724	0.670	0.979	0.982	
banana - Ori	800	2	0.180	0.428	0.318	0.442	
banana	800	2	0.180	0.428	0.318	0.442	
2d-10c	800	9	0.714	0.992	0.997	0.890	
aml28	800	5	0.680	1.000	0.996	0.996	

Table A3. Cont.

Dataset	<i>n</i>	<i>k</i>	TriHex	<i>k</i> -Means	HAC	GMM	Win
2d-20c-no0	800	20	0.564	0.976	0.973	0.882	
D31	800	31	0.472	0.909	0.880	0.900	
blobs	300	3	0.445	0.903	0.829	0.903	
DS-577	577	3	0.521	0.903	0.995	0.974	
2d-4c-no4	800	4	0.393	0.747	0.996	0.567	
R15	600	15	0.264	0.993	0.982	0.896	
chainlink	800	2	0.159	−0.000	−0.001	0.912	
DS-850	800	5	0.162	0.879	0.990	0.977	

References

- Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer: New York, NY, USA, 2009.
- Palla, G.; Derényi, I.; Farkas, I.; Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **2005**, *435*, 814–818.
- Lancichinetti, A.; Fortunato, S.; Kertész, J. Detecting the overlapping and hierarchical community structure in complex networks. *New J. Phys.* **2009**, *11*, 033015.
- Hackbusch, W. *Multi-Grid Methods and Applications*; Springer-Verlag: Berlin/Heidelberg, Germany, 1985.
- Ward, J.H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244.
- Johnson, S.C. Hierarchical clustering schemes. *Psychometrika* **1967**, *32*, 241–254.
- Batagelj, V.; Ferligoj, A.; Doreian, P. Direct and indirect methods for structural equivalence. *Soc. Networks* **1992**, *14*, 63–90.
- Murtagh, F.; Contreras, P. Algorithms for hierarchical clustering: An overview. *WIREs Data Min. Knowl. Discov.* **2012**, *2*, 86–97.
- Jiang, L.; Fang, H.; Shang, X.; Zhang, Z. Hierarchical clustering weighted Gaussian process regression low-fidelity model and application in airfoil drag coefficients prediction. *IEEE Trans. Aerosp. Electron. Syst.* **2026**, *62*, 9070–9081. <https://doi.org/10.1109/TAES.2026.3678593>.
- Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **1982**, *43*, 59–69.
- Kohonen, T. The self-organizing map. *Proc. IEEE* **1990**, *78*, 1464–1480.
- Kohonen, T. *Self-Organizing Maps*, 3rd ed.; Springer: Berlin/Heidelberg, Germany, 2001.
- Miikkulainen, R. Script recognition with hierarchical feature maps. *Connect. Sci.* **1990**, *2*, 83–101.
- Koikkalainen, P.; Oja, E. Self-organizing hierarchical feature maps. *Proc. IJCNN* **1990**, *2*, 279–284.
- Xie, J.; Szymanski, B.K.; Liu, X. SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In Proceedings of the IEEE ICDM Workshops, Vancouver, BC, Canada, 11 December 2011; pp. 344–349.
- Shi, J.; Malik, J. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 888–905.
- Ng, A.; Jordan, M.; Weiss, Y. On spectral clustering: Analysis and an algorithm. *Adv. Neural Inf. Process. Syst.* **2002**, *14*, 849–856.
- Grünbaum, B.; Shephard, G.C. *Tilings and Patterns*; W.H. Freeman: New York, NY, USA, 1987.
- Nadaraya, E.A. On estimating regression. *Theory Probab. Its Appl.* **1964**, *9*, 141–142.
- McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv* **2018**. arXiv:1802.03426.
- van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
- Halldórsson, M.M.; Radhakrishnan, J. Greedily achieving connectivity in directed graphs. *Scand. Workshop Algorithm Theory* **1994**, *805*, 201–212.
- Diday, E. *Une Représentation Visuelle des Classes Empiétantes: Les Pyramides*; Technical Report No. 291; INRIA: Le Chesnay-Rocquencourt, France, 1984.
- Diatta, J.; Fichet, B. From Apresjan hierarchies and Bandelt–Dress weak hierarchies to quasi-hierarchies. In *New Approaches in Classification and Data Analysis*; Diday E., Ed.; Studies in Classification, Data Analysis, and Knowledge Organization; Springer, Berlin/Heidelberg, Germany, 1994; pp.111–118. https://doi.org/10.1007/978-3-642-51175-2_11.
- Fichet, B. Sur une extension de la notion de hiérarchie et son équivalence avec quelques matrices de Robinson. *Actes Des Journées De Stat. De La Gd. Motte* **1984**, *12*.
- Bandelt, H.-J.; Dress, A.W.M. Weak hierarchies associated with similarity measures: An additive clustering technique. *Bull. Math. Biol.* **1989**, *51*, 133–166.
- Bandelt, H.-J.; Dress, A.W.M. A canonical decomposition theory for metrics on a finite set. *Adv. Math.* **1992**, *92*, 47–105.
- Diday, E. Orders and overlapping clusters by pyramids. In *Multidimensional Data Analysis*; de Leeuw, J., Ed.; DSWO Press: Leiden, The Netherlands, 1986; pp. 201–234.
- Bertrand, P.; Diday, E. A visual representation of the compatibility between an order and a dissimilarity index: The pyramids. *Comput. Stat. Q.* **1985**, *2*, 31–42.

30. Bertrand, P.; Diday, E. Une généralisation des arbres hiérarchiques: Les représentations pyramidales. *Rev. De Stat. Appl.* **1990**, *38*, 53–78.
31. Bertrand, P.; Janowitz, M.F. Pyramids and weak hierarchies in the ordinal model for clustering. *Discret. Appl. Math.* **2002**, *122*, 55–81.
32. Durand, C.; Fichet, B. One-to-one correspondences in pyramidal representation: A unified approach. In *Classification and Related Methods of Data Analysis*; Bock, H.H., Ed.; North-Holland: Amsterdam, The Netherlands, 1988; pp. 85–90.
33. Butler, S. Interlacing for weighted graphs using the normalized Laplacian. *Electron. J. Linear Algebra* **2006**, *16*, 90–98.
34. Bezdek, J.C. FCM: The fuzzy *c*-means clustering algorithm. *Comput. Geosci.* **1984**, *10*, 191–203.
35. Sibson, R. SLINK: An optimally efficient algorithm for the single-link cluster method. *Comput. J.* **1973**, *16*, 30–34.
36. Campello, R.J.G.B.; Moulavi, D.; Sander, J. Density-based clustering based on hierarchical density estimates. In *Proceedings of the PAKDD 2013, Gold Coast, Australia, 14–17 April 2013*; Lecture Notes in Computer Science; Springer: Berlin, Germany, 2013; Volume 7819, pp. 160–172.
37. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
38. Yang, J.; Leskovec, J. Overlapping community detection at scale: A nonnegative matrix factorization approach. In *Proceedings of the ACM WSDM, Rome, Italy, 4–8 February 2013*; pp. 587–596.
39. Lancichinetti, A.; Radicchi, F.; Ramasco, J.J.; Fortunato, S. Finding statistically significant communities in networks. *PLoS ONE* **2011**, *6*, e18961.
40. Newman, M.E.J.; Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E* **2004**, *69*, 026113.
41. Blondel, V.D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, *2008*, P10008.
42. Kenyon, R. The construction of self-similar tilings. *Geom. Funct. Anal.* **1996**, *6*, 471–488.
43. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**. *arXiv:1609.02907*.
44. Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S. How powerful are graph neural networks? *arXiv* **2018**. *arXiv:1810.00826*.
45. Hubert, L.; Arabie, P. Comparing partitions. *J. Classif.* **1985**, *2*, 193–218.
46. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65.
47. Caliński, T.; Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat.* **1974**, *3*, 1–27.
48. 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 Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, 2–4 August 1996*; pp. 226–231.
49. Clustering-Datasets: A Collection of Benchmark Datasets for Clustering Algorithms. GitHub Repository. 2021. Available online: <https://github.com/milaan9/Clustering-Datasets> (accessed on 24 July 2026).
50. Fränti, P.; Sieranoja, S. K-means properties on six clustering benchmark datasets. *Appl. Intell.* **2018**, *48*, 4743–4759. Available online: <http://cs.uef.fi/sipu/datasets/> (accessed on 24 July 2026).

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.