MASS-AWARE SPECTRAL CLUSTERING ON ADAPTIVE SUPERNODES

Anonymous authors
Paper under double-blind review

ABSTRACT

Spectral clustering can recover clusters of arbitrary shape, but only from the graph it is given, and graphs built from local affinities or compact representatives poorly describe extended clusters with unequal sample support. Density peaks provide complementary cues about which regions are well supported and well separated, but local prominence alone does not determine which regions share a label. We treat these cues as evidence for graph construction rather than as clustering decisions. The result is mass-aware spectral clustering (MASC) on adaptive supernodes: hierarchical grouping forms supernodes that follow the manifold structure and remain linked where they touch. Each nonroot supernode connects to its nearest predecessor in the mass order, forming a sparse peak-affinity tree whose weights are the reciprocal of mass times separation, so that heavy, prominent groups can be cut at low cost. A mass-normalized spectral objective then determines memberships jointly, and mass-weighted rounding propagates the same sample counts to the final labels. We prove that the reduced problem is exactly RatioCut on a lifted sample-level graph over supernode-respecting partitions and derive a boundary perturbation bound whose boundary-to-gap ratio predicts spectral stability. MASC tunes a single resolution parameter besides the cluster count and requires no kernel bandwidth or neighborhood size, and the eigenproblem scales with the number of supernodes rather than the number of samples. On real datasets, MASC attains the highest ACC and NMI against density-peak and spectral baselines, with average gains of 9.11/4.17 points over the best density-peak baseline and 9.62/5.79 over the best spectral baseline on each dataset.

1 INTRODUCTION

By partitioning a similarity graph, spectral clustering recovers clusters that need not be compact around a single center (Shi & Malik, 2000; Ng et al., 2002; von Luxburg, 2007). A global partition can use only the relationships the graph encodes. When the graph conflates local contact with common membership, a better solver does not repair the result.

Graph construction has accordingly received considerable attention: self-tuning affinities adapt to local scales, and representative graphs reduce the cost of the eigenproblem (Zelnik-Manor & Perona, 2004; Yan et al., 2009; Laplante & Ambroise, 2024). Recent work constructs coresets, combines anchor-based spectral models, and learns the graph jointly with the partition (Jourdan et al., 2025; Zhang et al., 2025; Kong et al., 2025). Two problems remain. First, local proximity graphs and compact representatives such as centroids or Voronoi cells describe a neighborhood or a ball. An extended class is neither. Its contact with a neighboring class often passes through a few close pairs at a single location. In a proximity graph that contact resembles any other short link, and a centroid or cell may erase that location or fragment the class. Second, a typical representative graph assigns every vertex the same weight. A small peripheral fragment and a region containing many samples thus enter the cut on equal terms, and the objective does not weigh a boundary against the population it divides. Both problems lie in the graph, not the solver, and raise the question of what a vertex should represent and what its edges should carry.

Density peaks supply the evidence of support and separation that proximity omits. Density-peak clustering (DPC) scores each observation by its local density and its distance to the nearest denser observation (Rodriguez & Laio, 2014; Li et al., 2025), thereby quantifying how well a region is

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

supported and how far it lies from a better-supported one. A cut requires both to place a boundary. DPC, however, treats prominent peaks as centers. In high-dimensional data one class often contains several local peaks, and adopting them as centers can split that class. Peak evidence should therefore propose boundaries rather than fix centers. We argue that the graph must carry the two things the cut will need: where groups touch and how many samples each holds.

MASC constructs such a graph with one design choice for each requirement. To preserve where groups touch, we form supernodes by hierarchical grouping with minimum-pair contact, so that a group follows the shape of a class rather than a ball around a center. The closest pair between two groups records where they touch, a contact that compact representatives can lose. To carry sample counts, we order supernodes by mass and connect each nonroot supernode to its nearest predecessor. The edge weight is the reciprocal of mass times separation, so that a heavy, prominent group is cheap to cut whereas a small fragment at ordinary separation remains attached. A mass-normalized spectral cut then assigns supernodes to clusters, and the rounding step weights each supernode by the same sample count. The sole parameter besides the cluster count is the resolution at which the hierarchy is cut.

Our main contributions are:

1. We propose MASC, a spectral clustering framework that converts density-peak evidence on adaptive supernodes into a sparse graph of candidate boundaries and carries sample mass through the partition.
2. The mechanism couples minimum-pair contact between adaptive groups, reciprocal edge weights derived from mass-times-separation saliency, and sample mass carried through cut normalization and rounding. Local peaks thus serve as boundary evidence rather than final centers, and graph reduction becomes a representation choice.
3. We establish an exact lifted RatioCut interpretation, an exact characterization of connected two-way cuts, and a boundary perturbation bound whose boundary-to-gap ratio predicts spectral stability in controlled experiments.

2 Related work

Affinity learning and spectral inference. Self-tuning spectral clustering adapts local bandwidths to heterogeneous scales (Zelnik-Manor & Perona, 2004). Joint graph learning and partitioning bring the affinity closer to the clustering objective (Kong et al., 2025); neural approaches learn cut assignments or spectral representations (He et al., 2025; Guo et al., 2026). Recent multi-view clustering also learns the graph with block-diagonal or connectivity structure (Zhao et al., 2026; Gu et al., 2026) or adopts a unified probabilistic perspective (Qin & Feng, 2026). Latent diffusion has been employed to denoise embeddings for clustering (Meier et al., 2026). These methods refine edge weights or the optimizer but leave open how an extended group and its population should be represented in the first place. MASC starts from this question: it represents groups as supernodes, derives their edges from support and separation, and uses the hierarchy only to select one resolution of supernodes, which keeps the graph sparse and the parameterization minimal.

Representative graphs and graph reduction. KASP/RASP approximate spectral clustering with representative centers (Yan et al., 2009), and Spectral Bridges derives affinities from samples within Voronoi cells (Laplante & Ambroise, 2024). FSEC combines anchor-based spectral models (Zhang et al., 2025), and prototypes serve the same role in incomplete multi-view clustering (Yu et al., 2025). Graph coarsening and coresets preserve properties of an input graph while reducing its size (Loukas, 2019; Jourdan et al., 2025), and weighted cuts (Dhillon et al., 2007) allow vertex weights to enter a cut objective. Compact representatives can lose contacts between groups, and preserving an input graph cannot repair poor initial geometry. MASC addresses both limitations: it retains group memberships and minimum-pair contact and carries exact masses through affinities, cut normalization, and rounding.

Density peaks and structural hierarchies. DPC and its variants refine density estimation, center selection, or membership assignment (Rodriguez & Laio, 2014; Hou et al., 2024; Zang et al., 2024; Ding et al., 2025). First-neighbor hierarchies (Sarfraz et al., 2019) and adjacency-constrained hierarchical clustering (Yang & Lin, 2024) provide related structural tools, and recent work studies

internal evaluation of density-based clusterings with noise (Beer et al., 2026). Improved local evidence, however, does not establish each prominent region as a cluster. MASC therefore adopts hierarchical grouping, uses density-peak evidence only to weight the peak-affinity tree, and determines memberships jointly through a mass-normalized spectral objective.

3 METHOD

Given observations $X = (x_r)_{r=1}^n$, $x_r \in \mathbb{R}^d$, and a cluster count $1 < K < n$, MASC tunes only the resolution $\lambda \in (0, 1]$. Dissimilarities $d(x_r, x_s)$ are assumed finite, symmetric, nonnegative, and zero on the diagonal; no triangle inequality is required. The algorithm aggregates sample indices into *supernodes* and partitions the resulting graph with supernodes weighted by sample count.

3.1 ADAPTIVE SUPERNODES FROM LOCAL CONTACT

We adopt adjacency-constrained hierarchical clustering (CHC; Yang & Lin, 2024) to form local groups without committing to final centers. Groups are compared by their minimum-pair distance, the single-linkage distance d_{sl}, which preserves the location at which two extended groups touch even when their remaining members are distant:

$$d_{\mathrm{sl}}(A, B) = \min_{r \in A,\, s \in B} d(x_r, x_s). \tag{1}$$

A *unit* is a set of sample indices treated as one object. Starting from the singleton units $\mathcal{C}^{(0)}$, each unit A selects its nearest distinct unit $\nu_h(A)$. CHC accepts the directed link $A \to \nu_h(A)$ only when $|A| \le |\nu_h(A)|$: a unit may join a neighbor of equal or larger mass but never links to a smaller one. Undirected connected components of accepted links constitute $\mathcal{C}^{(h+1)}$; groupwise minima update distances exactly. Contraction cannot stall, because a minimum-mass unit satisfies the rule against every other unit; the hierarchy thus reaches a single unit at depth $H \le n - 1$.

Write $M_h = |\mathcal{C}^{(h)}|$. The resolution selects a level, backing off when necessary so that units outnumber requested clusters:

$$h_\lambda = \min\{H, \max\{1, \lfloor \lambda H + \tfrac{1}{2} \rfloor\}\}, \qquad h^\star = \max\{h \in \{0, \dots, h_\lambda\} : M_h > K\}. \tag{2}$$

Since $M_0 = n > K$, the admissible set is nonempty. The selected supernodes $\{C_i\}_{i=1}^M = \mathcal{C}^{(h^\star)}$ have masses $q_i = |C_i|$ and $Q = \mathrm{diag}(q_1, \dots, q_M)$; only this level enters the graph. Mass counts the samples a supernode contains and quantifies structural support rather than an estimated density.

3.2 TURNING PEAK EVIDENCE INTO AFFINITY

The next stage converts support and separation into boundary costs, not into centers. Order the supernodes by decreasing mass, breaking ties by canonical order, and write $j \prec i$ for precedence; Appendix A, Proposition 4, treats signatures and permutation equivariance. Each nonroot supernode links to its nearest predecessor:

$$p(i) = \arg\min_{j \prec i} d_{\mathrm{sl}}(C_i, C_j), \qquad \delta_i = d_{\mathrm{sl}}(C_i, C_{p(i)}), \qquad \gamma_i = q_i \delta_i. \tag{3}$$

Ties in the $\arg\min$ are broken by canonical order. Because every link points earlier in the order, these $M - 1$ links form a tree rooted at the first supernode i_0, the *peak-affinity tree*. The *saliency* γ_i, the product of mass and separation by which DPC selects centers, serves here as evidence. A large saliency makes the edge to $p(i)$ cheap to cut, whereas a small fragment at ordinary separation remains attached. Proposition 2 quantifies the cost of removing an edge. We encode this evidence as reciprocal affinities:

$$W_{i,p(i)} = W_{p(i),i} = a_i = \frac{1}{\max\{\gamma_i, \tau_\gamma\}}, \qquad \tau_\gamma = \varepsilon_\gamma \min_{j \neq i_0 : \gamma_j > 0} \gamma_j. \tag{4}$$

Here $\varepsilon_\gamma = 10^{-6}$ is held fixed throughout and affects only zero-distance edges. All other affinities vanish; if every nonroot saliency is zero, we set $\tau_\gamma = 1$. When positive saliencies exist, scaling all distances by $\beta > 0$ scales W by β^{-1} and preserves the cut minimizers and generalized eigenspaces, so the method is invariant to the unit of distance. For edge $e = (i, p(i))$, write $a_e = a_i$, $\gamma_e = \gamma_i$, and $s_e = a_e^{-1}$.

Algorithm 1 MASC: mass-aware spectral clustering on adaptive supernodes

Require: Samples X, cluster count K, resolution λ

1: Construct the CHC hierarchy: accept each nearest-unit link whose source mass does not exceed its target mass, and merge connected components.
2: Select level $h^\star$ by Eq. (2); retain supernode memberships and masses q_i.
3: Order supernodes by decreasing mass, breaking ties by canonical order; construct the peak-affinity tree by Eq. (3) and assign affinities W by Eq. (4).
4: Solve $L_W u = \eta Q u$ for the bottom-K generalized eigenvectors, normalize their rows, and round by mass-weighted k-means with k-means++ initialization, R_{km} restarts, and K nonempty clusters.
5: **return** Sample labels $y \in [K]^n$, each sample inheriting its supernode's label.

3.3 A GLOBAL CUT WITH SAMPLE MASS

We determine memberships jointly, counting samples rather than supernodes. For a partition $\Pi = \{S_c\}_{c=1}^K$ into nonempty sets, let $q(S) = \sum_{i \in S} q_i$ and let $\mathrm{cut}_W(S, \bar{S})$ sum crossing affinities. Instantiating the weighted graph-cut framework (Dhillon et al., 2007) with sample mass yields the mass-normalized cut objective

$$\mathcal{J}(\Pi) = \sum_{c=1}^{K} \frac{\mathrm{cut}_W(S_c, \bar{S}_c)}{q(S_c)}. \tag{5}$$

Mass plays two roles: $q_i \delta_i$ determines the strength of one edge, whereas $q(S_c)$ balances an entire cluster. Since supernodes are uneven, treating a hundred-sample and a two-sample supernode as equal vertices would weight two samples as heavily as a hundred. Degree normalization addresses a different question: a vertex degree can be large because a few edges are strong rather than because a region contains many samples. Dividing by $q(S_c)$ instead charges a cut per separated sample, so each boundary is weighed against the population it divides. Writing $D_W = \mathrm{diag}(W\mathbf{1})$ and $L_W = D_W - W$, we relax partition indicators to obtain

$$\min_{U^\top Q U = I_K} \mathrm{tr}(U^\top L_W U), \qquad L_W u = \eta Q u. \tag{6}$$

With eigenvalues ordered increasingly, we row-normalize the bottom-K generalized eigenvectors to $\bar{U}$ and discretize them by mass-weighted rounding over surjective assignments $z : [M] \to [K]$:

$$\min_{z, \{\mu_c\}} \sum_{i=1}^{M} q_i \|\bar{u}_i - \mu_{z_i}\|_2^2, \qquad z([M]) = [K], \quad \mu_c \in \mathbb{R}^K. \tag{7}$$

We solve Eq. (7) by weighted k-means with k-means++ initialization over R_{km} restarts, retaining the assignment with the lowest weighted objective and K nonempty clusters. Supernodes are listed in canonical order throughout (Appendix A). Each sample inherits its supernode's label. Algorithm 1 summarizes the procedure.

4 THEORETICAL PROPERTIES

All statements in this section are proved in Appendix A.

Proposition 1 (Mass-consistent relaxation)**.** *Let $Z_{ic} = q(S_c)^{-1/2}$ for $i \in S_c$ and zero otherwise. Then $Z^\top Q Z = I_K$ and $\mathrm{tr}(Z^\top L_W Z) = \mathcal{J}(\Pi)$, so Eq. (6) is its Rayleigh–Ritz relaxation.*

The relaxation replaces the indicator matrix Z by any real U with $U^\top Q U = I_K$; the same Q appears because the indicators are normalized by cluster mass and are consequently Q-orthonormal.

The objective also admits an exact sample-level interpretation. Let $R \in \{0,1\}^{n \times M}$ encode membership, with $R^\top R = Q$. Distribute each supernode affinity uniformly over its sample pairs: $\widetilde{W} = RQ^{-1} W Q^{-1} R^\top$. With $\widetilde{D} = \mathrm{diag}(\widetilde{W}\mathbf{1})$ and $\widetilde{L} = \widetilde{D} - \widetilde{W}$, this gives

$$R^\top \widetilde{W} R = W, \qquad R^\top \widetilde{D} R = D_W, \qquad R^\top \widetilde{L} R = L_W. \tag{8}$$

Consequently $F = RU$ satisfies $F^\top F = U^\top Q U$, $F^\top \widetilde{L} F = U^\top L_W U$, and $F^\top \widetilde{D} F = U^\top D_W U$. The lifted cluster $\widetilde{S}_c = \bigcup_{i \in S_c} C_i$ has size $q(S_c)$ and the same crossing weight. Thus the objective is RatioCut on this lifted sample graph, restricted to supernode-respecting partitions and embeddings (von Luxburg, 2007), and mass-weighted rounding coincides with sample-level k-means on the

lifted row-normalized embedding with intact supernodes. Since the lifted graph is constructed from the supernode affinities, the identities in Eq. (8) hold exactly, not approximately. Replacing Q by D_W yields the restricted Ncut relaxation.

Proposition 2 (Exact connected two-way restriction)**.** *Let deleting tree edge e produce sides of sample masses m_e and $n - m_e$. Among these connected two-way partitions, the exact optimum is*

$$e^\star \in \arg\max_e s_e m_e (n - m_e). \tag{9}$$

Deleting e yields objective $\mathcal{J}(e) = n/[s_e m_e (n - m_e)]$, and an $O(M)$ postorder scan computes all subtree masses. For positive saliencies the score reduces to $\gamma_e m_e (n - m_e)$, which balances local separation against global mass. For $K > 2$, component masses depend jointly on all removed edges, so the objective no longer decomposes edge by edge. We therefore solve the K-way problem jointly through the relaxation of Eq. (6) for every K; Proposition 2 serves as a characterization, not as an algorithmic step.

Finally, we examine when weak boundaries leave the embedding stable. Deleting the target boundaries produces a forest of K components; stability depends on the strength of the deleted edges relative to the forest's internal spectral gap.

Theorem 3 (Boundary perturbation bound)**.** *Delete boundary set B to obtain exactly K tree components with Laplacian L_0. Define $L_B = L_W - L_0$, $A_0 = Q^{-1/2} L_0 Q^{-1/2}$, and the positive semidefinite update $E = Q^{-1/2} L_B Q^{-1/2}$. Let $0 = \eta_1(A_0) = \cdots = \eta_K(A_0) < \eta_{K+1}(A_0) =: \alpha$. For mass-normalized component indicators Z_0, the orthonormal columns of $V_0 = Q^{1/2} Z_0$ span $\mathcal{V}_0 = \ker(A_0)$. Let $\widehat{\mathcal{V}}$ be the bottom-K eigenspace of $\widehat{A} = A_0 + E$. If $\|E\|_2 < \alpha$, then*

$$\begin{aligned} \|\sin\Theta(\widehat{\mathcal{V}}, \mathcal{V}_0)\|_2 &\le \frac{\|EV_0\|_2}{\eta_{K+1}(\widehat{A})} \le \frac{\|Q^{-1/2} L_B Z_0\|_2}{\alpha} \le \frac{\|E\|_2}{\alpha}, \\ \|E\|_2 &\le \sum_{e=(i,j)\in B} \frac{1}{s_e} \left(\frac{1}{q_i} + \frac{1}{q_j} \right). \end{aligned} \tag{10}$$

Here Θ denotes principal angles. The residual Davis–Kahan argument (Davis & Kahan, 1970) measures the disturbance EV_0 to the ideal indicators; positive semidefiniteness gives $\eta_{K+1}(\widehat{A}) \ge \alpha$, and rank-one edge contributions yield the second bound. In words, V_0 is the ideal embedding of the K components, and the computed eigenspace remains close to it whenever the boundary edges are weak relative to the internal gap α. Each boundary edge contributes $1/s_e$ times $(1/q_i + 1/q_j)$, so a high-saliency boundary between heavy supernodes perturbs the embedding little. These are the edges that Eq. (4) makes weak. Section 5 confirms on controlled graphs that $\|E\|_2/\alpha$ tracks the eigenspace error. The bound concerns the eigenspace alone. If in addition the row-normalized embedding, after orthogonal alignment to the ideal embedding, has maximum row error below $\sqrt{2}/2$, assignment to the nearest ideal centroid recovers the component labels exactly (Appendix A).

Complexity. For Algorithm 1, assume an $O(d)$ sample-distance evaluation, R_{km} rounding restarts, and at most T_{km} Lloyd iterations. An exact all-pairs formulation admits the upper bound

$$O\big(n^2 d + n^2 \log n + M^3 + R_{\mathrm{km}} T_{\mathrm{km}} M K^2\big). \tag{11}$$

The terms cover pairwise distances, hierarchy construction and ordering, direct spectral solution, and rounding. At level h, groupwise minima and canonical ordering cost $O(M_h^2 \log M_h)$; since every minimum-mass unit merges while isolated units have larger integer mass, $M_h \le n/(h+1)$ and $\sum_h M_h^2 = O(n^2)$ (Appendix A). The $O(M^2)$ predecessor search and $O(n)$ label propagation are of lower order; storage is $O(nd + n^2)$. Indexed or approximate nearest-neighbor search, using k-d trees (Bentley, 1975), VP-trees (Yianilos, 1993), or locality-sensitive hashing (Datar et al., 2004), can reduce the neighbor-search cost further. With $O(d \log n)$ group-aware queries and $H = O(\log n)$ levels, the distance and hierarchy terms fall to roughly $O(nd \log^2 n)$.

5 Experiments

5.1 Experimental setup

Datasets. We evaluate MASC on ten real-world datasets, namely four face datasets, JAFFE (Lyons et al., 1998), UMIST (Graham & Allinson, 1998), CMU PIE (Sim et al., 2002), and YTF (Wolf et al.,

Table 1: ACC (top) and NMI (bottom) on ten datasets. Best result in bold, second best underlined.

Dataset	MASC	DPC (2014)	FDPAN (2025)	DBSCAN-DPC (2024)	DPC-DVND (2024)	NJW-SC (2002)	Spectral Bridges (2024)	Coreset-SC (2025)	NeuNcut (2025)	FSEC (2025)	BootSC (2026)
					ACC						
JAFFE	**.981**	.873	.765	.793	.915	.967	.831	.742	.643	.915	.798
miceprotein	**.554**	.526	.490	.525	.549	.551	.457	.474	.442	.535	.432
UMIST	**.845**	.616	.670	.581	.758	.619	.703	.428	.325	.619	.475
CMU PIE	**1.000**	.789	.310	.709	.367	.601	.450	.236	.053	.415	.230
MNIST	**.909**	.444	.499	.686	.510	.706	.780	.576	.452	.788	.659
COIL-100	**.914**	.548	.509	.619	.806	.809	.769	.534	.539	.751	.587
COIL-20	**.898**	.630	.708	.802	.872	.826	.867	.592	.448	.832	.658
YTF	**.679**	.635	.621	.614	.613	.658	.667	.440	.480	.657	.538
Libras Movement	**.536**	.458	.428	.336	.525	.519	.511	.389	.372	.511	.469
COIL-40	**.996**	.638	.727	.750	.866	.801	.878	.635	.585	.794	.666
					NMI						
JAFFE	**.974**	.933	.880	.916	.927	.964	.864	.736	.784	.927	.788
miceprotein	**.566**	.482	.488	.502	.489	.520	.440	.488	.383	.519	.406
UMIST	**.925**	.792	.789	.742	.881	.810	.841	.594	.531	.816	.649
CMU PIE	**1.000**	.953	.578	.891	.640	.763	.626	.456	.164	.686	.577
MNIST	**.838**	.533	.494	.731	.631	.749	.767	.583	.577	.710	.596
COIL-100	**.979**	.863	.826	.855	.945	.937	.947	.799	.780	.892	.837
COIL-20	**.963**	.795	.859	.890	.950	.890	.941	.764	.612	.930	.761
YTF	**.841**	.795	.781	.819	.803	.810	.830	.606	.662	.815	.745
Libras Movement	**.674**	.628	.593	.513	.666	.613	.622	.464	.548	.638	.580
COIL-40	**.996**	.883	.889	.893	.959	.925	.966	.802	.790	.892	.829

2011); three object-image datasets, COIL-20 (Nene et al., 1996a), COIL-40 (Nene et al., 1996b), and COIL-100 (Nene et al., 1996b); the handwritten-digit dataset MNIST (LeCun et al., 1998); and two tabular datasets, miceprotein (Higuera et al., 2015) and Libras Movement (Dias et al., 2009). Dataset statistics are given in Table 2 of Appendix C.

Baselines and experimental settings. We compare MASC with ten clustering methods: four density-peak methods, DPC (Rodriguez & Laio, 2014), FDPAN (Ding et al., 2025), DBSCAN-DPC (Hou et al., 2024), and DPC-DVND (Zang et al., 2024); and six spectral methods, NJW-SC (Ng et al., 2002), Spectral Bridges (Laplante & Ambroise, 2024), Coreset-SC (Jourdan et al., 2025), NeuNcut (He et al., 2025), FSEC (Zhang et al., 2025), and BootSC (Guo et al., 2026). The number of clusters is provided when required, and the parameters of all baselines are tuned over the ranges recommended in their original papers. Apart from K, MASC has only one free parameter, λ, which is searched over $\{0.05, 0.10, \ldots, 1.00\}$. For each setting, we run 20 weighted k-means++ restarts and report ACC and NMI.

5.2 Comparison with density-peak and spectral methods

Against the ten baselines in Table 1, MASC attains the highest ACC and NMI on all ten datasets. Relative to the best density-peak baseline on each dataset, its mean ACC/NMI margin is 9.11/4.17 percentage points, and relative to the best spectral baseline it is 9.62/5.79. The largest ACC margins arise on five image datasets: CMU PIE (1.000 against 0.789), MNIST (0.909 against 0.788), COIL-40 (0.996 against 0.878), COIL-100 (0.914 against 0.809), and UMIST (0.845 against 0.758). The margins narrow on the low-dimensional tabular datasets, miceprotein with 77 features (0.554 against 0.551) and Libras Movement with 90 features (0.536 against 0.525), and on YTF (0.679 against 0.667) and JAFFE (0.981 against 0.967), where several baselines already perform comparably. This pattern indicates that neighborhood and supernode graphs can encode comparable structure on relatively simple or low-dimensional data. On structured data such as images, the larger margins suggest that the supernode representation preserves gradual variation in viewpoint or pose better than neighborhood or centroid graphs.

COIL-100 and MNIST further illustrate this difference. In both datasets, samples of the same cluster vary substantially in viewpoint or writing style, so it is hard to preserve intra-class connectivity while keeping inter-class boundaries distinct. DPC may split clusters that contain several local density peaks, and DPC-DVND strengthens local evidence through voting and diffusion but still relies on sequential propagation. NJW-SC partitions globally from sample-level neighborhood relations, whereas Spectral Bridges connects compact Voronoi cells through directional affinities; such representations may be less suited to extended cluster structures with gradual intra-cluster variation. By contrast, MASC explicitly preserves the minimum-pair contacts between supernodes and weights each boundary according to the mass on both sides (Section 3).

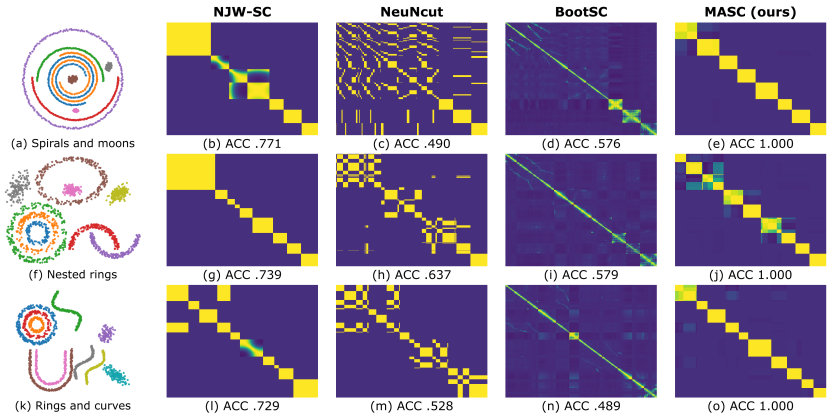

Figure 1: Three synthetic datasets and pairwise similarity matrices $\exp(-\|z_i - z_j\|^2)$ of the spectral embeddings produced by NJW-SC, NeuNcut, BootSC, and MASC. Samples are ordered by class; rows contain 8, 9, and 10 classes, and panel values report ACC.

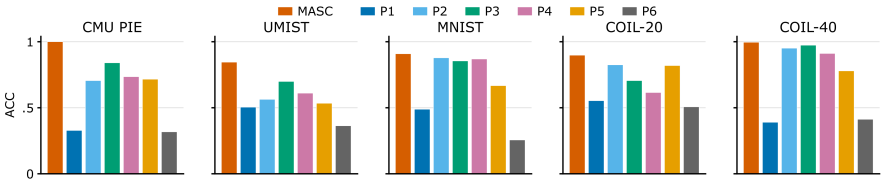

Figure 2: ACC of MASC and six ablation variants (P1–P6, defined in the text) on five datasets.

5.3 Qualitative analysis on synthetic shapes

Fig. 1 presents three synthetic datasets that probe how each method models structure. In these examples, the gap between two classes can be shorter than the extent of a single class, so that nearest neighbors may belong to different classes. No single center represents a ring or crescent adequately. In the similarity matrices, NJW-SC exhibits merged blocks and strong off-diagonal similarities, indicating class fusion. NeuNcut produces stripes and fragmented blocks, reflecting intra-class fragmentation. BootSC retains a bright local diagonal but displays weak global coherence within each class. In contrast, MASC forms well-defined class-level blocks and attains ACC 1.000 on all three datasets. Intra-class variation remains visible in the nested-rings example, yet every sample receives the correct label. We attribute this behavior mainly to two factors, both ablated on real data in Section 5.4: minimum-pair contacts between adaptive supernodes preserve connectivity across extended structures, and the mass-aware global cut over the peak-affinity tree places boundaries at contacts that are weak relative to the mass on both sides.

5.4 Ablations of grouping and graph construction

Fig. 2 compares MASC with six variants on five datasets. **P1** replaces the adaptive supernodes with average-linkage groups while retaining the same number M as MASC. **P2** replaces the minimum-pair distance with the mean-pair distance. **P3** and **P4** replace the peak-affinity tree with union kNN and ϵ-radius graphs, respectively, using locally scaled Gaussian weights. **P5** removes the grouping step and solves sample-level RatioCut on a union kNN graph with equal-weight rounding, and **P6** uses matched-count farthest-first representatives followed by nearest assignment. For P1 and P6, the peak-affinity and mass-aware spectral rules are retained after recomputing masses and contacts, whereas P3–P5 use fixed size-based graph settings.

Structural grouping is more than compression. MASC outperforms every variant, and replacing its grouping is the most damaging modification: on MNIST, P1/P6 reduce ACC from 0.909 to 0.489/0.256, and on CMU PIE from 1.000 to 0.328/0.318. With matched group counts and identical downstream rules, graph size cannot account for this gap; the memberships themselves determine which contacts and shared labels the representation can express. Neither all-pair proximity nor farthest-first coverage follows the geometry of an extended class, so the contacts required by the tree are lost before the spectral stage sees them.

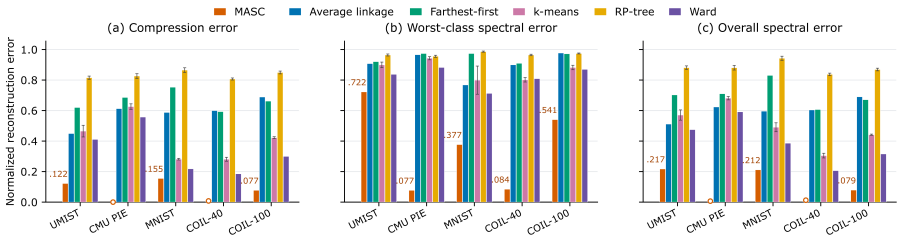

Figure 3: (a) Compression, (b) worst-class spectral, and (c) overall spectral error before rounding for MASC and five matched-count grouping controls that share MASC's affinity rule and spectral head. Lower is better; open circles mark bars too small to draw.

Contacts and peak evidence play distinct roles. P2 degrades performance only marginally on MNIST and COIL-40 but reduces ACC on UMIST from 0.845 to 0.563, showing that mean-pair distances obscure the close contacts between extended groups. Replacing the peak-affinity tree reduces ACC on COIL-20 from 0.898 to 0.706/0.615 for P3/P4, and P3 degrades slightly on COIL-40 (0.996 to 0.974). Support and separation therefore provide more reliable boundary evidence than proximity alone.

A sample-level spectral head is insufficient. P5 attains ACC 0.667 on MNIST and 0.534 on UMIST, compared with 0.909 and 0.845 for MASC, which shows that sample-count normalization cannot recover the structural advantage of MASC. Although P5 applies RatioCut to the sample graph, it lacks supernode grouping, minimum-pair contacts, and peak-informed affinities. Its spectral stage consequently remains driven by proximity and inherits the ambiguities observed in Section 5.3.

5.5 Does compression preserve the structure needed by spectral inference?

Equally small graphs need not be equally informative. To quantify how much class structure survives compression, we replace only the grouping step with average linkage, farthest-first assignment, Ward, k-means, or a random-projection (RP) tree, holding the supernode count M, affinity rule, and spectral head fixed. Only the grouping step changes, so these controls isolate its effect. The k-means and RP-tree results are averaged over five runs.

Following the projection view of graph reduction (Loukas, 2019), we measure three errors, defined in full in Appendix B. The *compression error* is the cluster structure lost when samples are merged into supernodes. The *overall spectral error* is the residual after projection onto the lifted bottom-K eigenspace. The *worst-class error* is the largest class-wise residual. Because the embedding lies in the span of the supernode indicators, the overall error cannot fall below the compression error.

As Fig. 3 shows, MASC preserves substantially more cluster structure than the alternative groupings. On CMU PIE, its compression error is zero and its overall error only 0.006, whereas every control incurs a compression error of at least 0.557. On COIL-40, the corresponding errors are 0.008 and 0.013, against 0.186 and 0.206 for Ward, the strongest control. Unlike centroid-based k-means, coverage-based farthest-first assignment, and projection-based RP-tree grouping, minimum-pair contact allows supernodes to follow extended class structures without enforcing compactness, so that contrasts between neighboring classes are preserved.

The remaining datasets exhibit the same trend. On MNIST, the worst-class error is 0.377 for MASC against 0.713 for Ward. On COIL-100, the compression error of 0.077 increases only to 0.079 after spectral truncation. On UMIST, the overall error is 0.217 for MASC and 0.475 for Ward. These results indicate that the advantage arises before rounding: informative supernodes retain class contrasts that the spectral head can exploit. Graph reduction is thus a representation choice, not merely compression. In a reduced graph, forming the vertices matters as much as weighting the edges.

5.6 Resolution, running time, and stability

Robustness to the resolution. The ACC curves in Fig. 4a exhibit broad, stable plateaus on five datasets, except for a dip on miceprotein around $\lambda \in [0.25, 0.35]$. At $\lambda = 0.45$, MASC attains its highest ACC on all five datasets, which suggests that a single setting suffices. The plateaus arise

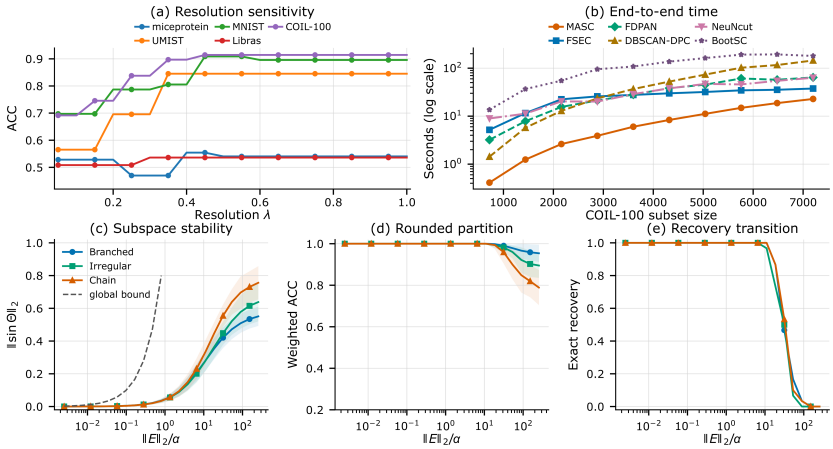

Figure 4: (a) ACC against the resolution λ. (b) Running time on COIL-100 subsets. (c–e) Eigenspace error, mass-weighted ACC, and exact recovery against $\|E\|_2/\alpha$ of Theorem 3. The dashed curve in (c) is the bound $\|E\|_2/\alpha$ itself; shading in (c–d) denotes one standard deviation across controlled graphs.

because λ selects a discrete hierarchy level, so that distinct values mapping to the same level yield identical supernodes. Smaller λ preserves finer structural detail but may retain stray fragments, whereas larger λ forms heavier supernodes at the risk of irreversible merges.

Running time. On COIL-100 subsets with n from 720 to 7,200 in steps of 720, MASC requires 0.412–22.84 seconds per end-to-end run and is faster than FSEC, FDPAN, DBSCAN-DPC, NeuNcut, and BootSC at all scales (Fig. 4b). After grouping, both the tree and the spectral problem operate only on supernodes, so their additional cost is relatively small.

Boundary strength is relative to internal cohesion. Theorem 3 bounds the eigenspace error by $\|E\|_2/\alpha$ whenever this ratio is below one, implying stability when boundary edges are weak relative to the internal spectral gap. We evaluate the bound on 90 branched, irregular, and chain graphs at 23 boundary-to-interior strength ratios. In Fig. 4c, the eigenspace error closely tracks the ratio (Spearman $\rho = 0.989$) and remains below the bound. Fig. 4d–e show that all 1,080 evaluations with $\|E\|_2/\alpha < 1$ attain ACC 1.000 and exact recovery, and degradation appears only at much larger ratios.

6 Conclusion

MASC is a mass-aware spectral clustering framework built on adaptive supernodes, and its central idea is that the graph must carry the two things the cut will need: where groups touch, and how many samples each holds. Adaptive supernodes linked by minimum-pair contact keep the place where extended groups touch without assuming compactness. A peak-affinity tree turns mass and separation into edge costs, so prominent groups are cheap to cut while small fragments stay attached. A global spectral partition that carries sample mass through cut normalization and rounding decides membership. Because the same masses appear throughout, the reduced problem is exactly RatioCut on a lifted sample-level graph over supernode-respecting partitions. The connected two-way optimum is available in closed form, and a boundary perturbation bound predicts when weak boundaries leave the embedding stable. On ten datasets, MASC attains the highest ACC and NMI against ten baselines. The ablations show that each of its three components is needed, and its supernodes retain class contrasts that matched-count groupings lose before the spectral step. Future work includes automatic selection of the resolution λ, multi-view contacts (Zhao et al., 2026; Qin & Feng, 2026), and group-aware indexing for further scaling.

AI USE STATEMENT

Generative AI tools were used only to assist with language polishing and improve the grammar and readability of the manuscript; the authors take full responsibility for all content.

REFERENCES

Anna Beer, Lena Krieger, Pascal Weber, Martin Ritzert, Ira Assent, and Claudia Plant. Internal evaluation of density-based clusterings with noise. In *The Fourteenth International Conference on Learning Representations*, pp. 86149–86172, 2026.

Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the Twentieth Annual Symposium on Computational Geometry*, pp. 253–262, 2004.

Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Inderjit S. Dhillon, Yuqiang Guan, and Brian Kulis. Weighted graph cuts without eigenvectors: A multilevel approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(11): 1944–1957, 2007.

Daniel Dias, Sarajane Peres, and Helton Bíscaro. Libras Movement. UCI Machine Learning Repository, 2009.

Shifei Ding, Chao Li, Xiao Xu, Lili Guo, Ling Ding, and Xindong Wu. Fast density peaks clustering algorithm based on approximate k-nearest neighbors. *IEEE Transactions on Knowledge and Data Engineering*, 37(10):5878–5889, 2025.

Daniel B. Graham and Nigel M. Allinson. Characterising virtual eigensignatures for general purpose face recognition. In *Face Recognition: From Theory to Applications*, pp. 446–456. Springer, 1998.

Zhibin Gu, Zhenhao Zhong, Xi Zhang, and Bing Li. Constant degree matrix-driven incomplete multi-view clustering via connectivity-structure and embedding tensor learning. In *The Fourteenth International Conference on Learning Representations*, pp. 84649–84671, 2026.

Wengang Guo, Wei Ye, Chunchun Chen, Xin Sun, Christian Böhm, Claudia Plant, and Susanto Rahardja. Bootstrap deep spectral clustering with optimal transport. *IEEE Transactions on Multimedia*, 28:531–544, 2026.

Wei He, Shangzhi Zhang, Chun-Guang Li, Xianbiao Qi, Rong Xiao, and Jun Guo. Neural normalized cut: A differential and generalizable approach for spectral clustering. *Pattern Recognition*, 164:111545, 2025.

Clara Higuera, Katheleen J. Gardiner, and Krzysztof J. Cios. Self-organizing feature maps identify proteins critical to learning in a mouse model of Down syndrome. *PLoS ONE*, 10(6):e0129126, 2015.

Jian Hou, Houshen Lin, Huaqiang Yuan, and Marcello Pelillo. Flexible density peak clustering for real-world data. *Pattern Recognition*, 156:110772, 2024.

Ben Jourdan, Gregory Schwartzman, Peter Macgregor, and He Sun. Coreset spectral clustering. In *The Thirteenth International Conference on Learning Representations*, pp. 53311–53326, 2025.

Lingyi Kong, Jingjing Xue, Feiping Nie, and Xuelong Li. Direct spectral clustering with new graph learning for better fitting. *IEEE Transactions on Knowledge and Data Engineering*, 37(7):3991–4002, 2025.

Félix Laplante and Christophe Ambroise. Spectral Bridges. *Computo*, 2024.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Chao Li, Shifei Ding, Xiao Xu, Lili Guo, Ling Ding, and Xindong Wu. Vertical federated density peaks clustering under nonlinear mapping. *IEEE Transactions on Knowledge and Data Engineering*, 37(2):1004–1017, 2025.

Andreas Loukas. Graph reduction with spectral and cut guarantees. *Journal of Machine Learning Research*, 20(116):1–42, 2019.

Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with Gabor wavelets. In *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 200–205, 1998.

Dominik Meier, Shixing Yu, Sagnik Nandy, Promit Ghosal, and Kyra Gan. Clustering by denoising: Latent plug-and-play diffusion for single-cell embeddings. In *The Fourteenth International Conference on Learning Representations*, pp. 19456–19481, 2026.

Sameer A. Nene, Shree K. Nayar, and Hiroshi Murase. Columbia Object Image Library (COIL-20). Technical Report CUCS-005-96, Columbia University, 1996a.

Sameer A. Nene, Shree K. Nayar, and Hiroshi Murase. Columbia Object Image Library (COIL-100). Technical Report CUCS-006-96, Columbia University, 1996b.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pp. 849–856, 2002.

Yalan Qin and Guorui Feng. Unified and efficient multi-view clustering from probabilistic perspective. In *The Fourteenth International Conference on Learning Representations*, pp. 101262–101274, 2026.

Alex Rodriguez and Alessandro Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, 2014.

Saquib Sarfraz, Vivek Sharma, and Rainer Stiefelhagen. Efficient parameter-free clustering using first neighbor relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8934–8943, 2019.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

Terence Sim, Simon Baker, and Maan Bsat. The CMU pose, illumination, and expression (PIE) database. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 53–58, 2002.

Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 529–534, 2011.

Donghui Yan, Ling Huang, and Michael I. Jordan. Fast approximate spectral clustering. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 907–916, 2009.

Jie Yang and Chin-Teng Lin. Enhanced adjacency-constrained hierarchical clustering using fine-grained pseudo labels. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8 (3):2481–2492, 2024.

Peter N. Yianilos. Data structures and algorithms for nearest neighbor search in general metric spaces. In *Proceedings of the Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 311–321, 1993.

Shengju Yu, Zhibin Dong, Siwei Wang, Pei Zhang, Yi Zhang, Xinwang Liu, Naiyang Guan, Tiejun Li, and Yiu-ming Cheung. Simple yet effective incomplete multi-view clustering: Similarity-level imputation and intra-view hybrid-group prototype construction. In *The Thirteenth International Conference on Learning Representations*, pp. 6971–6997, 2025.

Wenke Zang, Jing Che, Linlin Ma, Xincheng Liu, Aoyu Song, Jingwen Xiong, Yuzhen Zhao, Xiyu Liu, Yawen Chen, and Hui Li. Density peaks clustering based on density voting and neighborhood diffusion. *Information Sciences*, 681:121209, 2024.

Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems*, pp. 1601–1608, 2004.

Runxin Zhang, Shuaijun Hang, Zhensheng Sun, Feiping Nie, Rong Wang, and Xuelong Li. Anchor-based fast spectral ensemble clustering. *Information Fusion*, 113:102587, 2025.

Shuping Zhao, Yulong Chen, Jie Wen, Lunke Fei, Jinrong Cui, and Tingting Chai. Confident block diagonal structure-aware invariable graph completion for incomplete multi-view clustering. In *The Fourteenth International Conference on Learning Representations*, pp. 99860–99870, 2026.

A PROOFS

A.1 LEVEL-COUNT BOUND USED IN THE COMPLEXITY ANALYSIS

Let m_h be the minimum unit mass at level h, so $m_0 = 1$. A minimum-mass unit A satisfies $|A| \le |\nu_h(A)|$ and so merges, producing a unit of mass at least $2m_h \ge m_h + 1$, whereas a unit A' that remains isolated has its link rejected, so $|A'| > |\nu_h(A')| \ge m_h$. Hence $m_{h+1} \ge m_h + 1$, $m_h \ge h + 1$, $M_h \le n/m_h \le n/(h+1)$, and $\sum_h M_h^2 \le n^2 \sum_{h\ge 0}(h+1)^{-2} = O(n^2)$.

A.2 PROOF OF PROPOSITION 1

Write $Z_{:,c}$ for column c of Z. Disjoint supports give $Z_{:,c}^\top Q Z_{:,c'} = \mathbf{1}\{c = c'\}$, and the Laplacian identity

$$Z_{:,c}^\top L_W Z_{:,c} = \tfrac{1}{2}\sum_{i,j} W_{ij}(Z_{ic} - Z_{jc})^2 = \frac{\mathrm{cut}_W(S_c, \bar{S}_c)}{q(S_c)}$$

holds because only edges crossing between S_c and $\bar{S}_c$ contribute. Summing over c proves $Z^\top Q Z = I_K$ and the discrete objective identity. For the relaxation, $V = Q^{1/2}U$ gives $V^\top V = I_K$ and

$$\mathrm{tr}(U^\top L_W U) = \mathrm{tr}\Big(V^\top Q^{-1/2} L_W Q^{-1/2} V\Big).$$

By Rayleigh–Ritz, this trace minimization is solved by the bottom-K eigenvectors of the symmetric operator; transforming back with $U = Q^{-1/2}V$ yields $L_W u = \eta Q u$. The positive-weight tree is connected, so $\eta_1 = 0 < \eta_2$ and every bottom-K Q-orthonormal basis U spans a subspace containing the constant vector $\mathbf{1}$. The row norms $\|u_i\|_2 = (UU^\top)_{ii}^{1/2}$ are invariant under $U \mapsto UO$ for orthogonal O and therefore equal those of a basis containing the column $\mathbf{1}/\sqrt{n}$, so $\|u_i\|_2 \ge 1/\sqrt{n} > 0$ and the row normalization is well-defined.

A.3 PROOF OF THE LIFTING IDENTITIES (EQ. (8))

The lifted graph $\widetilde{W} = RQ^{-1}WQ^{-1}R^\top$ distributes each supernode affinity uniformly over sample pairs: a pair with $r \in C_i$ and $s \in C_j$ receives weight $W_{ij}/(q_i q_j)$, and no intra-supernode edge arises because $W_{ii} = 0$. The membership matrix satisfies $R^\top R = Q$, and

$$(R^\top \widetilde{W} R)_{ij} = \sum_{r\in C_i}\sum_{s\in C_j} \frac{W_{ij}}{q_i q_j} = W_{ij}.$$

Every sample in C_i has lifted degree $q_i^{-1}\sum_j W_{ij}$; summing over its q_i samples gives $(R^\top \widetilde{D} R)_{ii} = \sum_j W_{ij}$, while the off-diagonal entries vanish because $\widetilde{D}$ is diagonal and each row of R has a single nonzero entry. Hence $R^\top \widetilde{D} R = D_W$, and subtracting yields $R^\top \widetilde{L} R = L_W$, which establishes Eq. (8). Substituting $F = RU$ proves the three quadratic-form identities of the main text, $F^\top F = U^\top Q U$, $F^\top \widetilde{L} F = U^\top L_W U$, and $F^\top \widetilde{D} F = U^\top D_W U$.

For $\widetilde{S}_c = \bigcup_{i\in S_c} C_i$, $|\widetilde{S}_c| = q(S_c)$ and summing its crossing weights gives $\mathrm{cut}_{\widetilde{W}}(\widetilde{S}_c, \bar{\widetilde{S}}_c) = \mathrm{cut}_W(S_c, \bar{S}_c)$, so $\mathcal{J}(\Pi)$ is the RatioCut of $\{\widetilde{S}_c\}$ on the lifted graph. For the row-normalized lift $\bar{F} = R\bar{U}$ and labels inherited from supernodes, each supernode contributes q_i identical rows, so

$$\sum_{r=1}^{n} \|\bar{f}_r - \mu_{y_r}\|_2^2 = \sum_{i=1}^{M} q_i \|\bar{u}_i - \mu_{z_i}\|_2^2,$$

where $y_r = z_i$ for $r \in C_i$. This proves the rounding-objective equivalence. Finally, $\mathrm{vol}_{\widetilde{W}}(\widetilde{S}_c) = \sum_{i\in S_c}(D_W)_{ii}$ establishes the restricted-Ncut interpretation.

A.4 PROOF OF PROPOSITION 2

Throughout, $\mathcal{J}(e)$ and $\mathcal{J}(B)$ denote the objective of the partition obtained by deleting edge e or edge set B from the tree. Deleting edge e leaves connected sides of masses m_e and $n - m_e$ with e

as the only crossing edge, so

$$\mathcal{J}(e) = a_e \left(\frac{1}{m_e} + \frac{1}{n - m_e} \right) = \frac{n}{s_e m_e (n - m_e)}.$$

Minimizing this positive expression over the tree edges yields Eq. (9). For a multi-edge deletion B, let $S_i(B)$ denote the component containing supernode i. Regrouping the cluster boundary terms by edge gives

$$\mathcal{J}(B) = \sum_{e=(i,j)\in B} a_e \left(\frac{1}{q(S_i(B))} + \frac{1}{q(S_j(B))} \right).$$

Each denominator is the mass of a tree component after deletion of the entire set B, so an edge's contribution depends on which other edges are deleted, and the K-way optimum is in general not attained by selecting the $K-1$ largest two-way scores of Eq. (9). For instance, on a four-vertex path with unit masses, unit affinities, and $K = 3$, the two-way scores of the three edges are $3, 4, 3$, yet deleting the two end edges gives $\mathcal{J} = 1 + 1 + 1 = 3$, whereas any pair containing the middle edge gives $\mathcal{J} = 3.5$. By Eq. (4), $s_e = \gamma_e$ for positive nonroot saliencies, and the fixed floor gives $s_e = \tau_\gamma$ for zero-saliency edges. Rescaling all weights by a common positive factor leaves the minimizing edge unchanged, since it multiplies the objective, the Laplacian eigenvalues, and the boundary perturbation by the same factor and preserves the eigenspaces and every ratio in Eq. (10).

A.5 Proof of Theorem 3

Write $L_e = a_e(\boldsymbol{\varepsilon}_i - \boldsymbol{\varepsilon}_j)(\boldsymbol{\varepsilon}_i - \boldsymbol{\varepsilon}_j)^\top$ for the Laplacian of the single edge $e = (i, j)$, where $\boldsymbol{\varepsilon}_i$ is coordinate vector i; then L_W is the sum of L_e over all tree edges, L_0 the sum over the retained edges, and $L_B = L_W - L_0 = \sum_{e\in B} L_e$, so $E = \sum_{e\in B} Q^{-1/2} L_e Q^{-1/2}$ is positive semidefinite. Deleting B leaves exactly K tree components, so $\ker(L_0)$ is spanned by the mass-normalized component indicators Z_0, $A_0 = Q^{-1/2} L_0 Q^{-1/2}$ has the K-dimensional kernel $Q^{1/2}\ker(L_0)$, and $\alpha = \eta_{K+1}(A_0)$ is its smallest nonzero eigenvalue. The columns of $V_0 = Q^{1/2} Z_0$ are orthonormal and span $\ker(A_0)$. Since $A_0 V_0 = 0$, $\widehat{A} V_0 = E V_0 = Q^{-1/2} L_B Z_0$. Let $\widehat{V}_\perp$ contain the eigenvectors of $\widehat{A}$ above its bottom-K subspace, with diagonal eigenvalue matrix $\widehat{\Lambda}_\perp$. Then

$$\widehat{\Lambda}_\perp \widehat{V}_\perp^\top V_0 = \widehat{V}_\perp^\top E V_0.$$

Since $\|\widehat{V}_\perp^\top V_0\|_2 = \|\sin\Theta(\widehat{\mathcal{V}}, \mathcal{V}_0)\|_2$, multiplying by $\widehat{\Lambda}_\perp^{-1}$ and taking operator norms yields the residual inequality in Eq. (10). Because E is positive semidefinite, Weyl monotonicity gives $\eta_{K+1}(\widehat{A}) \geq \alpha$; min–max also gives $\eta_K(\widehat{A}) \leq \|E\|_2 < \alpha$, so the bottom-$K$ truncation has a strict gap. Together with $E V_0 = Q^{-1/2} L_B Z_0$ and $\|E V_0\|_2 \leq \|E\|_2$, this proves the first line of the bound.

The normalized contribution of boundary edge $e = (i, j)$ is

$$Q^{-1/2} L_e Q^{-1/2} = a_e v_e v_e^\top, \qquad v_e = \boldsymbol{\varepsilon}_i / \sqrt{q_i} - \boldsymbol{\varepsilon}_j / \sqrt{q_j},$$

whose norm $a_e \|v_e\|_2^2 = s_e^{-1}(q_i^{-1} + q_j^{-1})$ is small for a weak edge between heavy supernodes. The triangle inequality over the boundary edges proves the second line.

The bound controls the eigenspace; the passage to labels relies on a rowwise-separation condition. In the ideal embedding Z_0, row normalization maps each component to a coordinate vector, so the ideal centroids are separated by $\sqrt{2}$. Let O be the orthogonal $K \times K$ matrix that best aligns the computed basis with V_0, and define the maximum row error as the largest Euclidean distance between a row of the aligned row-normalized embedding and its ideal coordinate vector. Whenever this maximum row error is below $\sqrt{2}/2$, assignment to the nearest ideal centroid recovers the component labels exactly.

A.6 Canonical ordering and permutation equivariance

At each hierarchy level, exact ties are broken by the lexicographic signatures

$$\sigma_i = (-q_i, \pi_i, \kappa(C_i)), \tag{12}$$

where π_i sorts all other-unit distances decreasingly and κ is the fixed feature fingerprint defined below. Signature ranks resolve tied choices and order the supernodes before seeded rounding; strict comparisons are unaffected.

The fingerprint uses four fixed projections of each sample, with coordinate-j weights $\sin(j\sqrt{2})$, $\cos(j\sqrt{3})$, $\sin(j\sqrt{5}+.5)$, and $\cos(j\sqrt{7}+.25)$. The rank of a sample is one plus the number of samples whose projection vector is lexicographically smaller, so equal projections share a rank and ranks depend only on values. The value $\kappa(C)$ is the minimum rank among the samples in C and is consulted only when the preceding mass and complete distance-profile fields agree exactly.

Proposition 4 (Permutation equivariance)**.** *Assume distinct signatures at every exact tie, a positive gap between eigenvalues K and $K+1$ of $Q^{-1/2}L_WQ^{-1/2}$, exact comparisons, and deterministic arithmetic that preserves permutation identities and resolves subsequent choices in an invariant way. With fixed-seed rounding after canonical ordering, input permutations preserve the partition up to label names.*

Proof. Let P permute the samples and let $\Delta = [d(x_r, x_s)]_{r,s}$ denote the dissimilarity matrix, which becomes $P\Delta P^\top$, while supernode masses and the sorted distance profiles in Eq. (12) are unchanged. The fingerprint $\kappa(C)$ depends only on the feature vectors in C, so it is invariant to sample order, and the distinct-signature assumption makes every required rank unique. Induction over hierarchy levels yields the same contracted sample sets. Since masses, minimum-pair distances, and fingerprints are unchanged, the signatures and hence the precedence order $\prec$ are the same; breaking ties in the $\arg\min$ of Eq. (3) by canonical order gives the same predecessors $p(i)$, saliencies γ_i, and floor τ_γ, so for the induced supernode permutation P_U,

$$Q' = P_U Q P_U^\top, \qquad L'_W = P_U L_W P_U^\top.$$

The positive Kth eigengap determines the selected generalized eigenspace, which is row-permuted and may undergo an orthogonal basis change $U \mapsto UO$. Row normalization commutes with that change, because $\|u_i O\|_2 = \|u_i\|_2$, so the normalized rows rotate by O and all Euclidean distances between rows and mass-weighted centroids used by rounding are unchanged. Canonical supernode ordering then yields the same fixed-seed weighted-k-means++ draws and Lloyd choices after undoing the permutation, so the partition changes only by label names. □

B Definitions of the reconstruction errors

Let $T = RQ^{-1/2}$, let V contain the bottom-K orthonormal eigenvectors of $Q^{-1/2}L_WQ^{-1/2}$, and let $F = TV$. Since $V = Q^{1/2}U$, this coincides with the lifted embedding $F = RU$ of Section 4. Let $\mathcal{G}_1, \dots, \mathcal{G}_K \subseteq [n]$ be the reference classes with $n_c = |\mathcal{G}_c|$, and write $G = [g_1, \dots, g_K]$ with $g_c = \mathbf{1}_{\mathcal{G}_c}/\sqrt{n_c}$. With $\mathbf{u}_0 = \mathbf{1}/\sqrt{n}$ and $\omega_c = n_c/n$, let $\widetilde{g}_c = (g_c - \sqrt{\omega_c}\,\mathbf{u}_0)/\sqrt{1-\omega_c}$. The errors in Fig. 3 are

$$\begin{aligned} E_{\text{comp}} &= \frac{\|(I - TT^\top)G\|_F^2}{K-1}, \qquad E_{\text{overall}} = \frac{\|(I - FF^\top)G\|_F^2}{K-1}, \\ E_{\text{worst}} &= \max_c \|(I - FF^\top)\widetilde{g}_c\|_2^2. \end{aligned} \tag{13}$$

Since T and F have orthonormal columns, $TT^\top$ and $FF^\top$ are orthogonal projectors. The class-indicator space, the supernode space, and the connected graph's bottom-K spectral space all contain the constant sample direction $\mathbf{u}_0$. Removing this uninformative direction leaves $K-1$ reference contrasts, so all three errors in Eq. (13) lie in $[0,1]$. Since $\operatorname{span}(F) \subseteq \operatorname{span}(T)$, the overall error equals the compression error plus the nonnegative loss incurred by spectral truncation:

$$E_{\text{overall}} - E_{\text{comp}} = \frac{\|(TT^\top - FF^\top)G\|_F^2}{K-1} \geq 0.$$

The overall error also equals the mean squared sine of the $K-1$ nonconstant principal angles between $\operatorname{span}(F)$ and the reference class space. With $\varrho_c = \|(I - FF^\top)\widetilde{g}_c\|_2^2$,

$$E_{\text{overall}} = \frac{\sum_c (1-\omega_c)\varrho_c}{K-1}, \qquad E_{\text{worst}} = \max_c \varrho_c,$$

so the two errors are, respectively, a weighted average and the maximum of the per-class residuals; for balanced classes the weights $(1-\omega_c)/(K-1)$ are uniform, recovering the plain average. Reference labels enter only through these errors, after grouping and eigendecomposition, and play no role in the grouping, affinity, or spectral steps that the errors evaluate. All controls share the per-dataset supernode count M used by MASC.

C DATASET STATISTICS

Table 2 lists the ten real datasets used in Section 5.

Table 2: Real datasets: number of samples (n), features (d), and classes (K).

Dataset	n	d	K	Dataset	n	d	K
JAFFE	213	1,024	10	COIL-100	7,200	49,152	100
miceprotein	1,077	77	8	COIL-20	1,440	16,384	20
UMIST	575	10,304	20	YTF	10,056	9,075	40
CMU PIE	2,856	1,024	68	Libras Movement	360	90	15
MNIST	10,000	784	10	COIL-40	2,880	49,152	40