STRUCTURE-GUIDED TRANSFORMER CLUSTERING FOR HIGH-DIMENSIONAL DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Clustering high-dimensional data is challenging when informative signals are distributed across dependent features and cluster structure is nonlinear. We introduce Structure-Guided Transformer Clustering (SGTC), which estimates a feature conditional-dependence graph, partitions related variables into groups, guides the splitting of Transformer heads, and learns a prototype-based latent partition. Across simulations, SGTC achieved higher agreement with ground-truth clusters than an unstructured Transformer, K-means, spectral clustering, and UMAP-based clustering, particularly as dimensionality increased. Its advantage was strongest when the sample size was sufficient to estimate useful dependence structure and diminished in sample-limited settings. SGTC showed robust performance under random edge deletions and additions to the feature-dependence graphs. We further applied SGTC to miRNA expression profiles from 418 patients with lung squamous cell carcinoma in The Cancer Genome Atlas. Among the evaluated methods, the clusters identified by SGTC showed the strongest overall survival separation (log-rank p=0.0066). The pathway enrichment analysis of the top driven miRNAs suggested more plausible pathways based on SGTC than the classic transformer.

1 INTRODUCTION

Unsupervised clustering is widely used to identify latent subgroups from high-dimensional data, including patient subgroups defined by molecular profiles. The task is especially challenging in common omics studies when the number of variables far exceeds the sample size, cluster structure is nonlinear, and informative signals are distributed across dependent features. Under these conditions, distance-based clustering can be dominated by redundant variables and may fail to recover the underlying cluster structure.

Deep clustering addresses nonlinear structure by learning a representation jointly with, or before, cluster assignment (Min et al., 2018). However, most methods learn the organization of the input features entirely from the same observations used for clustering. This can be inefficient for high dimensional features. Self-attention provides a flexible mechanism for learning nonlinear interactions among variables (Vaswani et al., 2017). An unstructured multi-head mechanism uses the natural/permuted order of input features, leaving a grouping space for borrowing feature-dependency information.

We consider an estimated feature-dependence graph as a structural inductive bias (Battaglia et al., 2018). By assigning related variables to the same Transformer heads, the graph can reduce the burden of learning feature organization from limited data. This guidance is not uniformly beneficial: the graph is also estimated from the observations and may be unreliable when the sample size is too small. A useful method should therefore utilize dependence structure and remain reasonably stable to graph-estimation error.

The TCGA lung squamous cell carcinoma (LUSC) cohort provides a biologically relevant application for structure-guided clustering. Previous mRNA analyses identified 4 reproducible LUSC subtypes: classical, basal, secretory, and primitive. Independently derived miRNA clusters were found to associate with these subtypes (Wilkerson et al., 2010; Cancer Genome Atlas Research Network, 2012). We focus on miRNA expression because neighboring and intronic miRNAs are often co-transcribed and co-expressed, and their expression profiles can reflect tumor lineage and differentiation state (Baskerville and Bartel, 2005; Lu et al., 2005). These dependencies make TCGA-LUSC

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.

a suitable setting for evaluating whether SGTC can identify biologically interpretable patient groups that complement, rather than necessarily reproduce, the established mRNA taxonomy.

We propose Structure-Guided Transformer Clustering (SGTC), a graph-guided multi-head encoder for unsupervised clustering. SGTC estimates conditional dependence among variables, partitions the resulting graph into related feature groups, guides Transformer head splitting, and learns a prototype-based latent partition. We consider both binary and weighted graph estimates and compare them with an unstructured Transformer and classical or manifold-learning clustering methods.

Our contributions are:

- We introduce a Transformer clustering framework that uses an estimated feature-dependence graph to guide head specialization without requiring a prespecified biological network.
- We evaluate SGTC under increasing dimensionality, different sample sizes, and graph perturbations. We then apply it to TCGA-LUSC miRNA data, using survival and pathway analyses for exploratory characterization.

2 Related Work

Representation-based deep clustering. Autoencoder-based methods combine representation learning with a clustering loss. DEC sharpens soft assignments through self-training (Xie et al., 2016); IDEC adds a reconstruction objective to preserve local structure (Guo et al., 2017); and DCN jointly optimizes an autoencoder and a K-means objective (Yang et al., 2017). These approaches balance two competing goals: reconstruction can retain variation unrelated to clustering, whereas self-training can reinforce inaccurate initial assignments. VaDE places a Gaussian-mixture prior on a variational latent representation (Jiang et al., 2017), making performance dependent on the suitability of the latent mixture assumption. DeepCluster alternates feature learning with pseudo-label assignment (Caron et al., 2018), but was developed for large image datasets with architectures and augmentations that do not transfer directly to modest-sample molecular data.

Graph-based and attention-based clustering. Graph-based deep clustering methods incorporate structural information, but usually represent observations or entities as graph nodes. DAEGC combines graph attention with self-training (Wang et al., 2019); SDCN combines an autoencoder with graph convolution over an observation-level neighborhood graph (Bo et al., 2020). SGTC addresses a different structure: its nodes are input variables, and its edges estimate conditional dependence among those variables. The graph determines how features are allocated across attention heads before patients are clustered. This distinction is important for molecular matrices, where dependence occurs among measured features even when no patient-level graph is available.

3 Method

3.1 Problem setup and overview

Let $\mathbf{X} = (x_{ij}) \in \mathbb{R}^{n \times m}$ denote the standardized data matrix for n subjects and m variables. We use $h \in \{1, \ldots, H\}$ for graph-guided feature groups and Transformer heads and $k \in \{1, \ldots, K\}$ for subject clusters. Thus, H controls the internal feature partition, whereas K is the number of output clusters.

Structure-Guided Transformer Clustering (SGTC) proceeds in four stages. It estimates a feature-dependence graph, partitions related variables into groups, constructs a graph-informed input for each Transformer head, and learns a latent representation using a prototype-based clustering objective. We consider a binary graph, SGTC-A, and a weighted graph, SGTC-Σ.

3.2 FEATURE GRAPH AND GRAPH-GUIDED GROUPS

Let $\widehat{\Sigma}_{\mathrm{LW}}$ be the Ledoit-Wolf shrinkage covariance estimate (Ledoit and Wolf, 2004). We define the estimated pseudo-precision matrix and edge scores as

$$\widehat{\Omega} = \widehat{\Sigma}_{\mathrm{LW}}^{+}, \qquad W_{j\ell} = |\widehat{\Omega}_{j\ell}|, \qquad W_{jj} = 0,$$

where $(\cdot)^+$ is the Moore-Penrose pseudoinverse. Under a multivariate Gaussian model, a zero off-diagonal precision entry indicates conditional independence (Friedman et al., 2008). We use $W_{j\ell}$ as an estimated conditional-dependence score, without treating thresholded sample edges as known conditional-independence relations. Denote the binary and weighted adjacency matrices as

$$A^{(A)}_{j\ell} = \mathbb{I}(|W_{j\ell}| > 0), \qquad A^{(\Sigma)}_{j\ell} = W_{j\ell},$$

and the Laplace matrix as $L = D - A$ where $D_{jj} = \sum_{\ell=1}^{m} A_{j\ell}$. Ledoit-Wolf shrinkage provides a stable covariance estimate when $m > n$; other precision estimators can be substituted without changing the remaining architecture.

Let $0 \le \lambda_1 \le \cdots \le \lambda_m$ be the eigenvalues of $L^{(v)}$. We select the number of groups over a prespecified candidate set $\mathcal{H}_{\mathrm{cand}}$ by

$$\widehat{H} = \arg \max_{h \in \mathcal{H}_{\mathrm{cand}}} (\lambda_{h+1} - \lambda_h).$$

For the remaining notation, $H = \widehat{H}$. We apply balanced constrained K-means to the rows of the first H Laplacian eigenvectors to obtain the feature partition $\mathcal{V} = \{\mathcal{V}_1, \ldots, \mathcal{V}_H\}$, with group sizes differing by at most one (Bradley et al., 2000). This constraint prevents very small or empty heads and balances computation; it does not assume equally sized biological modules. Zero-valued padding is added only when equal computational widths are required. For very large m, an optional anchor approximation partitions a subset of variables and assigns each remaining variable to its most strongly correlated anchor group.

3.3 GRAPH-INFORMED HEAD INPUTS

For subject i and group $\mathcal{V}_h$, let $b_{ih} \in \mathbb{R}^4$ contain the groupwise mean, variance, mean square, and mean absolute value of $\{x_{ij} : j \in \mathcal{V}_h\}$. The variance uses divisor $|\mathcal{V}_h|$. Let $\bar{b}_{ih} = S_b(b_{ih})$ denote componentwise standardization across subjects.

Principal component analysis is applied separately to $\mathbf{X}$. Let $p_i \in \mathbb{R}^{d_{\mathrm{PCA}}}$ be the retained scores for subject i. After zero-padding when needed, we split the scores into H blocks, $p_i = (p_i^{(1)}, \ldots, p_i^{(H)})$, and give head h the input

$$r_i^{(h)} = \begin{bmatrix} p_i^{(h)} \\ \alpha \bar{b}_{ih} \end{bmatrix}, \qquad \alpha \ge 0.$$

Each head therefore receives one graph-group summary and a share of the global principal-component representation. The entries of $r_i^{(h)}$ are embedded as tokens and processed by head-specific query, key, and value projections. Head outputs are concatenated and passed through an output projection and feed-forward layers with GELU activation, layer normalization, and dropout. The encoder returns $z_i = f_\theta(r_i) \in \mathbb{R}^d$, where $r_i = (r_i^{(1)}, \ldots, r_i^{(H)})$.

3.4 CLUSTERING

Let $\{c_k\}_{k=1}^K$ be learnable latent prototypes. We normalize the embeddings and prototypes as $\widetilde{z}_i = z_i / \|z_i\|_2$ and $\widetilde{c}_k = c_k / \|c_k\|_2$, and define

$$\ell_{ik} = \frac{\widetilde{z}_i^\top \widetilde{c}_k}{\tau}, \qquad p_{ik} = \frac{\exp(\ell_{ik})}{\sum_{k'=1}^{K} \exp(\ell_{ik'})},$$

where $\tau > 0$ is the temperature. Initial pseudo-labels are obtained by applying K-means to $\{r_i\}_{i=1}^n$. For current label y_i and $q_{ik} = \mathbb{I}(k = y_i)$, the mini-batch loss is

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \sum_{k=1}^{K} q_{ik} \log p_{ik}.$$

The encoder and prototypes are optimized jointly, and pseudo-labels are periodically refreshed by applying K-means to the current normalized embeddings. Complete optimization settings are given in Appendix Table 3.

3.5 THEORETICAL PROPERTIES

We characterize invariance of the group summaries, the geometry of prototype assignment, and label stability under feature-group error. These results are conditional on fitted representations and do not imply global convergence of the nonconvex training procedure.

Proposition 1 (permutation invariance). For any subject i and group $\mathcal{V}_h$, both b_{ih} and $\bar{b}_{ih}$ are invariant to permutations of the variables within $\mathcal{V}_h$.

Proposition 2 (nearest-prototype assignment). For normalized embeddings and prototypes,

$$\arg\max_k p_{ik} = \arg\max_k \widetilde{z}_i^\top \widetilde{c}_k = \arg\min_k \|\widetilde{z}_i - \widetilde{c}_k\|_2^2.$$

Thus, the hard assignment is independent of $\tau > 0$.

Theorem 1 (feature-group stability). Let $\mathcal{V}^* = \{\mathcal{V}_1^*, \dots, \mathcal{V}_H^*\}$ be a reference partition and $\widehat{\mathcal{V}}$ an estimated partition. After aligning groups, suppose the two versions of group h have size s_h and exchange r_h variables. Assume $|x_{ij}| \le M$; S_b is fixed and L_b-Lipschitz; the reference encoder f_{θ^*} is L_f-Lipschitz; and the relevant encoder outputs have norm at least $a > 0$. Further suppose

$$\sup_r \|f_{\widehat{\theta}}(r) - f_{\theta^*}(r)\|_2 \le \epsilon_\theta, \qquad \max_k \|\widehat{\widetilde{c}}_k - \widetilde{c}_k^*\|_2 \le \epsilon_c.$$

For reference labels y_i, define the cosine margin

$$\kappa = \min_i \left\{ (\widetilde{z}_i^*)^\top \widetilde{c}_{y_i}^* - \max_{k \ne y_i} (\widetilde{z}_i^*)^\top \widetilde{c}_k^* \right\} > 0,$$

and let

$$C_M = \sqrt{8M^2 + 40M^4},$$
$$\delta_{\mathcal{V}} = \left\{ \sum_{h=1}^{H} \left(\frac{r_h}{s_h} \right)^2 \right\}^{1/2},$$
$$\epsilon_z = \frac{2}{a} \left(\epsilon_\theta + L_f \alpha L_b C_M \delta_{\mathcal{V}} \right).$$

If $\kappa > 2(\epsilon_z + \epsilon_c)$, replacing $\mathcal{V}^*$ with $\widehat{\mathcal{V}}$ does not change any hard prototype assignment.

Corollary 1 (stability to graph perturbation). Let L^* be a reference Laplacian with eigengap $\gamma = \lambda_{H+1}(L^*) - \lambda_H(L^*) > 0$, and let $\widehat{L} = L^* + E$. Davis-Kahan perturbation theory gives

$$\min_{Q^\top Q = I} \|\widehat{U} Q - U^*\|_F \le c_{\mathrm{DK}} \frac{\|E\|_2}{\gamma}.$$

If the right-hand side is below the stability radius δ_G of the balanced spectral partition, then $\widehat{\mathcal{V}} = \mathcal{V}^*$. The prototype labels are then unchanged when $\kappa > 2(2\epsilon_\theta/a + \epsilon_c)$.

Corollary 2 (stability of the final K-means partition). Let μ_k^* and $\widehat{\mu}_k$ be aligned reference and perturbed K-means centroids of the normalized embeddings, with $\max_k \|\widehat{\mu}_k - \mu_k^*\|_2 \le \epsilon_\mu$. Define

$$\Delta = \min_i \min_{k \ne y_i} \left\{ \|\widetilde{z}_i^* - \mu_k^*\|_2^2 - \|\widetilde{z}_i^* - \mu_{y_i}^*\|_2^2 \right\} > 0.$$

If $\Delta > 8(\epsilon_z + \epsilon_\mu)$, all final K-means assignments are unchanged. See Appendix A for proofs.

Together, these results separate three sources of stability: the graph eigengap controls the feature partition, the encoder Lipschitz constant controls the propagation of representation error, and the clustering margin controls the sensitivity of subject assignments. They do not guarantee that the estimated graph is correct or that SGTC outperforms an unstructured method.

4 SIMULATIONS

4.1 DATA-GENERATING MECHANISM

We generated data according to 3 known clusters. Each cluster had a distinct block-diagonal covariance matrix containing 50 equal-sized blocks. Within each block, 20% of the possible off-diagonal entries were sampled without replacement and set to $\rho = 0.5$; the remaining off-diagonal entries were 0s, and the diagonal entries were 1. Edge locations were generated independently across clusters. Each block was adjusted to be positive semidefinite before observations were generated.

Cluster means were sparse. For each cluster, $\lfloor 0.25n \rfloor$ of the m variables were independently selected to carry nonzero mean signals.

For each observation, we first sampled a Gaussian vector using its cluster-specific mean and covariance matrix. To introduce nonlinear effects, the variables were divided into four equal subsets and transformed using $\sin(\pi x^2)$, $\cos(\pi x)$, $\sin(\pi x)$, and x^2, respectively. The final data retained 75% of the transformed component and 25% of the original Gaussian component. Feature columns were then globally permuted to remove the artificial contiguous ordering of the covariance blocks.

4.2 COMPARISON METHODS

We compared SGTC-A and SGTC-Σ with 4 reference methods: an unstructured Transformer (TC), K-means, spectral clustering, and UMAP followed by K-means. All methods were supplied with the true number of clusters, $K = 3$. The unstructured Transformer used the same prototype objective and final clustering procedure as SGTC but omitted the graph-guided representation. K-means was applied to the standardized features; spectral clustering used a 10-nearest-neighbor affinity graph; and UMAP reduced the standardized data to 20 dimensions using 50 neighbors and a minimum distance of zero before applying K-means. These experiments compare complete method configurations rather than provide a comprehensive ranking of deep clustering architectures. Complete implementation settings are provided in Appendix Table **??**. Each configuration included 50 independently generated datasets.

4.3 GRIDS AND GRAPH PERTURBATIONS

We considered sample sizes ($n \in 200, 500, 1000$), dimension-to-sample-size ratios ($m/n \in 1, 5, 10$), dependence levels ($\rho \in 0.3, 0.5, 0.7$), and sparsity levels ($s \in 0.1, 0.2, 0.4$). Each configuration was evaluated using 50 replicates.

We evaluated sensitivity to graph misspecification at ($n = 500, \rho = 0.5, s = 0.2$, and $m/n \in 1, 5, 10$). Within each replicate, we either removed 10% or 20% of the retained edges or added spurious edges equal to 10% or 20% of the retained edge count. Each 10% perturbation was nested within its corresponding 20% perturbation. Both SGTC variants received the same modified edge set, and added edges in SGTC-Σ were assigned the median weight of the retained precision edges.

4.4 EVALUATION METRICS

We evaluated agreement with the ground-truth labels using the Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), Fowlkes-Mallows Index (FMI), and Purity. Each metric has a maximum of one, with higher values indicating greater agreement. We report the mean and standard deviation across 50 replicates.

5 SIMULATION RESULTS

5.1 DIFFERENT DIMENSIONS

Figure 1 summarizes the focused dimension-stress benchmark. At $m/n = 1$, most methods were near ceiling. At $m/n = 5$, SGTC-A and SGTC-Σ achieved mean ARI of 0.973 and 0.974, mean NMI of 0.953 and 0.956, mean FMI of 0.982 and 0.983, and mean Purity of 0.991 for both. UMAP was the strongest non-graph comparator in ARI at 0.543. At $m/n = 10$, SGTC-A and SGTC-Σ

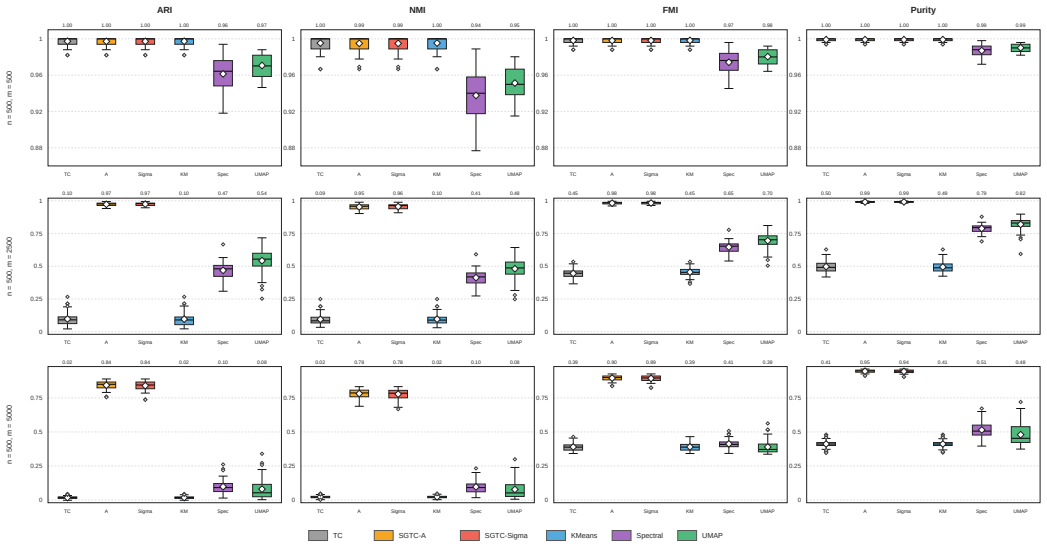

Figure 1: Dimension-stress comparison at $n = 500$, $\rho = 0.5$, and $s = 0.2$ over 50 paired replicates per setting. Rows correspond to $m/n \in \{1, 5, 10\}$ and columns to ARI, NMI, FMI, and Purity. Each panel compares TC, SGTC-A, SGTC-Σ, K-means, spectral clustering, and UMAP.

Table 1: Mean (SD) clustering agreement at $n = 500$, $m = 2500$, $\rho = 0.5$, and $s = 0.2$ over 50 paired replicates.

Method	ARI	NMI	FMI	Purity
TC	0.097 (0.052)	0.095 (0.043)	0.445 (0.036)	0.497 (0.044)
SGTC-A	0.973 (0.014)	0.953 (0.024)	0.982 (0.010)	0.991 (0.005)
SGTC-Σ	0.974 (0.014)	0.956 (0.022)	0.983 (0.009)	0.991 (0.005)
K-means	0.096 (0.052)	0.096 (0.043)	0.455 (0.035)	0.495 (0.044)
Spectral	0.469 (0.065)	0.413 (0.056)	0.646 (0.043)	0.788 (0.033)
UMAP	0.543 (0.095)	0.481 (0.083)	0.695 (0.063)	0.820 (0.053)

retained mean ARI of 0.843 and 0.840, whereas spectral clustering, the strongest non-graph comparator, reached 0.098. The two SGTC graph constructions performed similarly throughout the focused comparison.

5.2 Extended operating range

The effect of dimensional stress depended strongly on sample size. At $n = 200$, both SGTC variants had the highest mean ARI at all three ratios, but their mean ARI decreased from 0.758 at $m/n = 1$ to approximately 0.10 at $m/n = 5$ and 0.03 at $m/n = 10$. Thus, graph guidance did not overcome the sample-limited regime. The focused pattern was recovered at $n = 500$; at $n = 1000$, both variants reached mean ARI of 1.000 at $m/n \in \{1, 5\}$ and 0.994 at $m/n = 10$.

Across the nine dependence settings at $n = 500$ and $m/n = 5$, mean ARI for the SGTC variants ranged from 0.890 to 0.984. Agreement decreased as ρ and the nonzero off-diagonal proportion s jointly increased and was lowest at $\rho = 0.7$ and $s = 0.4$. Both variants nevertheless exceeded the comparison methods throughout this grid.

5.3 Robustness to graph misspecification

The SGTC variants were stable on average under the evaluated random graph perturbations. For every metric, variant, dimension ratio, and perturbation, the absolute value of the mean paired change from the corresponding unperturbed result was below 0.008. Replicate-level changes were less uni-

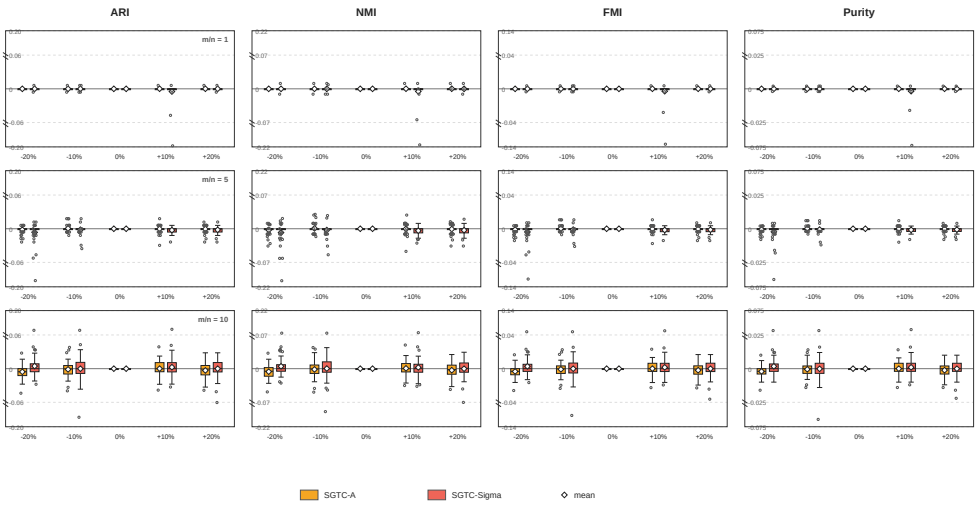

Figure 2: Graph-misspecification robustness at $n = 500$, $\rho = 0.5$, and $s = 0.2$ over 50 paired replicates per setting. Rows correspond to $m/n \in \{1, 5, 10\}$ and columns to ARI, NMI, FMI, and Purity. This figure supports average stability under tested random perturbations, but does not replace the graph-source ablation.

form, particularly after edge removal in higher dimensions. For SGTC-Σ, the most negative ARI changes were -0.164 under 20% edge removal at $m/n = 5$ and -0.146 under 10% edge removal at $m/n = 10$; some replicates instead improved. These findings support average robustness to the tested random edge perturbations, but not invariance for every dataset.

5.4 Computational cost

Across the simulation settings, mean runtimes for the two SGTC variants ranged from 1.3-2.7 seconds at $n = 200$, 3.5-5.2 seconds at $n = 500$, and 13.1-14.6 seconds at $n = 1000$ on the tested machine. Runtime increased approximately linearly with the number of variables, m. Appendix **??** describes the measurement protocol and reports results for each setting.

6 Application: TCGA-LUSC miRNA Clustering

6.1 Study Cohort and Analysis Method

We applied SGTC and other 4 comparators to identify molecular subgroups from miRNA expression profiles in The Cancer Genome Atlas lung squamous cell carcinoma (TCGA-LUSC) cohort (Cancer Genome Atlas Research Network, 2012). The analysis included 418 patients and 808 miRNAs. Because ground-truth cluster labels were unavailable, we evaluated post-diagnosis overall survival across the resulting patient groups using Kaplan-Meier curves and log-rank tests. Pairwise cluster comparisons were performed using Cox proportional hazards models adjusted for age at diagnosis and sex.

We also used Integrated Gradients (IG) to identify the top 50 miRNAs contributing most strongly to each partition Sundararajan et al. (2017). Precursor features were mapped to mature miRNAs and then to validated target genes. We then tested whether these target genes were over-represented in 2,868 curated pathways from Reactome (Fabregat et al., 2018; Milacic et al., 2024) for SGTC-Σ and TC, using hypergeometric tests (Reimand et al., 2019). Enrichment p-values were corrected for multiple testing using the Benjamini-Hochberg false discovery rate (FDR) Benjamini and Hochberg (1995). The primary analysis used the Reactome annotated-gene universe as the background. A sensitivity analysis used strongly validated targets reachable from any assayed precursor feature.

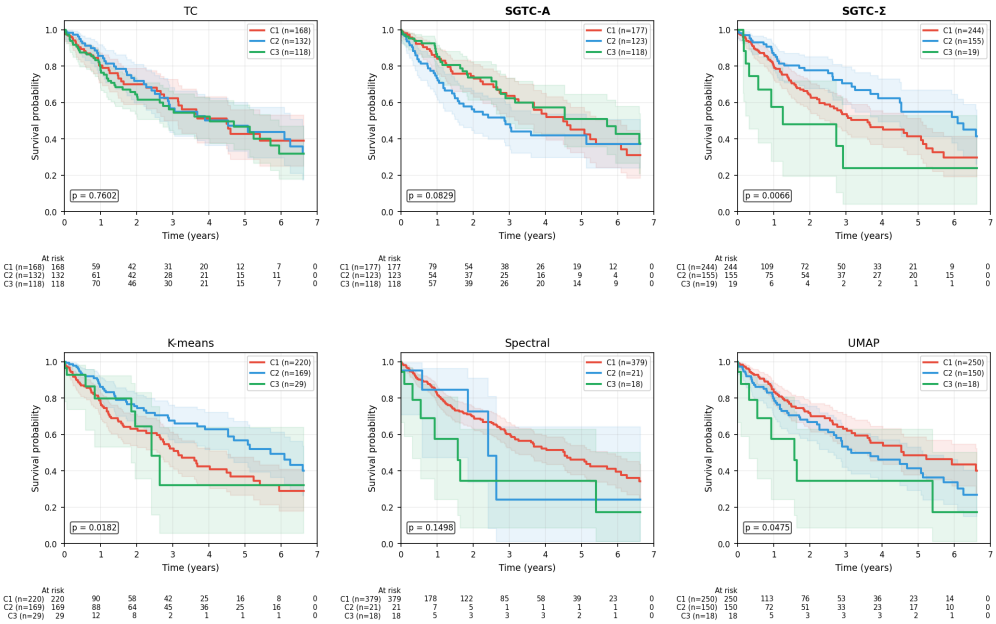

Figure 3: Kaplan-Meier curves for the three-cluster partitions identified from TCGA-LUSC miRNA expression ($n = 418$), with follow-up capped at 6.6 years.

Table 2: Survival distributions for the TCGA-LUSC three-cluster partitions ($n = 418$).

Method	Deaths/All			Cox Hazard Ratio (p-value)		Log-rank p
	C1	C2	C3	C2 vs C1	C3 vs C1	
TC	58/206	56/191	9/21	0.965 (0.851)	1.635 (0.171)	0.3272
SGTC-A	47/177	45/123	31/118	1.438 (0.082)	0.897 (0.638)	0.0829
SGTC-Σ	78/244	36/155	9/19	0.624 (0.020)	1.763 (0.108)	0.0066
K-means	71/220	44/169	8/29	0.589 (0.006)	0.997 (0.994)	0.0182
Spectral	110/379	5/21	8/18	1.242 (0.636)	1.991 (0.061)	0.1498
UMAP	68/251	46/148	9/19	1.112 (0.579)	2.204 (0.026)	0.0753

6.2 Survival Characterization of Identified Clusters

Figure 3 and Table 2 compared the survival distributions of 3 identified patient clusters from the 6 methods. The SGTC-Σ partition showed the strongest overall association with survival (log-rank p=0.0066). Compared with C1, C2 contained 155 patients and had a lower estimated hazard (HR=0.624, p=0.020); C3 contained only 19 patients and had a less precise, nonsignificant estimate (HR=1.763,p=0.108).

The K-means partition also showed an overall survival association (log-rank p=0.0182), with a lower estimated hazard for C2 than C1 (HR=0.589, p=0.006). Overall evidence was weaker for SGTC-A (p=0.083) and UMAP (p=0.075), with no significant difference for TC (p=0.327) or spectral clustering (p=0.150). The spectral partition was highly imbalanced, with 379 (91%) assigned to C1.

6.3 Pathway Enrichment Result

Among the 50 top-ranked SGTC-Σ precursor features, 13 mapped to 18 mature miRNAs and 210 unique validated targets, of which 182 were included in the Reactome annotated-gene universe. For TC, 21 features mapped to 29 mature miRNAs and 312 validated targets, of which 260 were included in Reactome. At FDR ≤ 0.05, the primary analysis identified 305 enriched terms for SGTC-Σ and 411 for TC. The sensitivity analysis identified 30 and 52 terms, respectively.

Both methods identified interleukin and PI3K/AKT signaling. Additional SGTC-Σ enrichments included receptor tyrosine kinase, MAPK-family, TGF-β, and VEGFA-VEGFR2 signaling, whereas TC identified broader signal-transduction and transcriptional programs, including RUNX, TP53, and insulin-receptor pathways. Figure 4 summarizes the enrichment results.

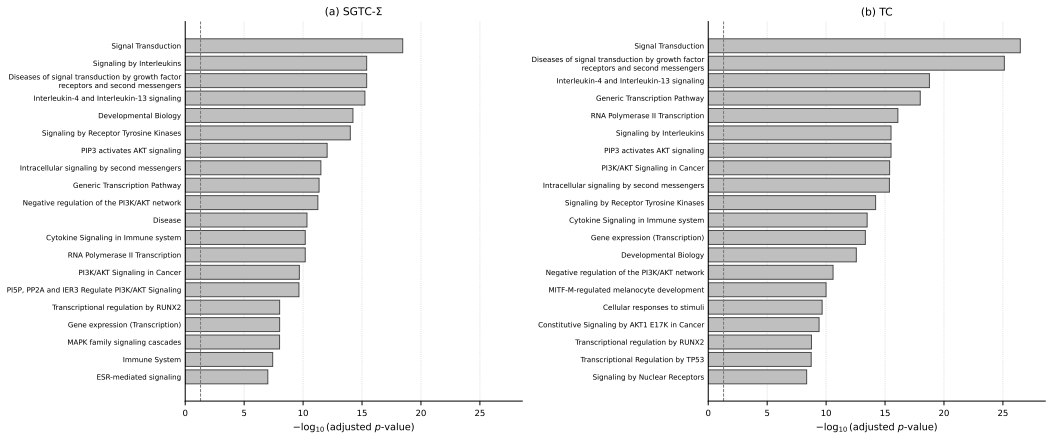

Figure 4: Reactome enrichment comparison for validated targets of the 50 highest-ranked features from the archived SGTC-Σ partition and an unstructured TC comparator.

7 Discussion

SGTC uses an estimated feature-dependence graph as an inductive bias for high-dimensional clustering. In simulations, both SGTC variants performed near the ceiling at (m/n=5) and retained high clustering agreement at (m/n=10), whereas the unstructured Transformer and standard clustering methods degraded substantially. However, performance declined at (n=200) and improved at (n=500) and (n=1000), indicating that SGTC depends on having sufficient data to estimate useful feature structure.

Random edge additions and deletions produced small average changes in clustering performance, although some individual replicates were more sensitive. These findings support average stability under the tested perturbations, not robustness to arbitrary graph misspecification. A further limitation is that the current comparisons do not isolate whether meaningful graph structure, rather than feature grouping or additional summary features, drives the improvement. A graph-source ablation comparing absent, random, true, and estimated graphs is therefore needed.

The TCGA-LUSC analysis illustrates a potential biological application of SGTC. The SGTC-Σ partition showed the strongest observed overall survival separation among the evaluated partitions, and attribution and Reactome enrichment analyses linked its top-ranked miRNAs to biologically relevant pathways.

8 Conclusion

SGTC is promising when feature dependence can be estimated reliably. It outperformed the comparison methods as dimensionality increased and was stable under the tested graph perturbations. Its advantage may be smaller when the sample size is insufficient to estimate the feature structure reliably.

AI Use Statement

ChatGPT was used to improve language and organization and revise Latex format. Claude was used to organize and implement algorithms and methods.

REFERENCES

Scott Baskerville and David P. Bartel. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA*, 11(3):241–247, 2005. doi: 10.1261/rna.7240905.

Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George E. Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018. doi: 10.48550/arXiv.1806.01261. URL https://arxiv.org/abs/1806.01261.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

Deyu Bo, Xiao Wang, Chuan Shi, Meiqi Zhu, Emiao Lu, and Peng Cui. Structural deep clustering network. In *Proceedings of The Web Conference 2020*, pages 1400–1410, 2020. doi: 10.1145/3366423.3380214. URL https://doi.org/10.1145/3366423.3380214.

Paul S. Bradley, Kristin P. Bennett, and Ayhan Demiriz. Constrained k-means clustering. Technical Report MSR-TR-2000-65, Microsoft Research, 2000.

Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519–525, 2012. doi: 10.1038/nature11404.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018.

Antonio Fabregat et al. Reactome: a curated knowledgebase of biological pathways. *Nucleic Acids Research*, 46(D1):D649–D655, 2018.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008. doi: 10.1093/biostatistics/kxm045.

Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. Improved deep embedded clustering with local structure preservation. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1753–1759, 2017.

Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: An unsupervised and generative approach to clustering. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 1965–1972, 2017. doi: 10.24963/ijcai.2017/273. URL https://doi.org/10.24963/ijcai.2017/273.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03)00096-4.

Jun Lu, Gad Getz, Eric A. Miska, Ezequiel Alvarez-Saavedra, Justin Lamb, David Peck, Alejandro Sweet-Cordero, Benjamin L. Ebert, Raymond H. Mak, Adolfo A. Ferrando, James R. Downing, Tyler Jacks, H. Robert Horvitz, and Todd R. Golub. MicroRNA expression profiles classify human cancers. *Nature*, 435(7043):834–838, 2005. doi: 10.1038/nature03702.

Marija Milacic et al. The reactome pathway knowledgebase 2024. *Nucleic Acids Research*, 52(D1):D672–D678, 2024. doi: 10.1093/nar/gkad1025. URL https://doi.org/10.1093/nar/gkad1025.

E. Min, X. Guo, Q. Liu, G. Zhang, J. Cui, and J. Long. A survey of clustering with deep learning: From the perspective of network architecture. *IEEE Access*, 6:39501–39514, 2018.

Jüri Reimand, Ruth Isserlin, Veronique Voisin, Mike Kucera, Christian Tannus-Lopes, Asha Rostamianfar, Lina Wadi, Mona Meyer, Jeff Wong, Changjiang Xu, Daniele Merico, and Gary D. Bader. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature Protocols*, 14(2):482–517, 2019. doi: 10.1038/s41596-018-0103-9. URL https://doi.org/10.1038/s41596-018-0103-9.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017. URL https://papers.nips.cc/paper/7181-attention-is-all-you-need.

Chun Wang, Shirui Pan, Ruiqi Hu, Guodong Long, Jing Jiang, and Chengqi Zhang. Attributed graph clustering: A deep attentional embedding approach. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 3670–3676, 2019. doi: 10.24963/ijcai.2019/509. URL https://doi.org/10.24963/ijcai.2019/509.

Matthew D. Wilkerson, Xiaoying Yin, Katherine A. Hoadley, Yufeng Liu, Michele C. Hayward, Christopher R. Cabanski, Kenneth Muldrew, C. Ryan Miller, Scott H. Randell, Mark A. Socinski, Alden M. Parsons, William K. Funkhouser, Carrie B. Lee, Patrick J. Roberts, Leigh Thorne, Philip S. Bernard, Charles M. Perou, and D. Neil Hayes. Lung squamous cell carcinoma mrna expression subtypes are reproducible, clinically important, and correspond to normal cell types. *Clinical Cancer Research*, 16(19):4864–4875, 2010. doi: 10.1158/1078-0432.CCR-10-0199.

J. Xie, R. Girshick, and A. Farhadi. Unsupervised deep embedding for clustering analysis. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

Bo Yang, Xiao Fu, Nicholas D. Sidiropoulos, and Mingyi Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 3861–3870, 2017.

A PROOFS FOR THE THEORETICAL PROPERTIES

The proofs below concern the clustering map conditional on fitted representations, not global convergence of alternating neural optimization.

A.1 PROOF OF PROPOSITION 1

Fix subject i, group $\mathcal{V}_h$, and a permutation $\pi : \mathcal{V}_h \to \mathcal{V}_h$. Each coordinate of

$$b_{ih} = \left(\frac{1}{|\mathcal{V}_h|} \sum_{j \in \mathcal{V}_h} x_{ij}, \frac{1}{|\mathcal{V}_h|} \sum_{j \in \mathcal{V}_h} (x_{ij} - \mu_{ih})^2, \frac{1}{|\mathcal{V}_h|} \sum_{j \in \mathcal{V}_h} x_{ij}^2, \frac{1}{|\mathcal{V}_h|} \sum_{j \in \mathcal{V}_h} |x_{ij}| \right)^\top$$

depends only on the multiset of values. Thus replacing j by $\pi(j)$ leaves all four coordinates unchanged: $b_{ih}(x_{i,\mathcal{V}_h}) = b_{ih}(x_{i,\pi(\mathcal{V}_h)})$. Componentwise standardization S_b follows the summaries, so $\bar{b}_{ih} = S_b(b_{ih})$ is invariant as well. □

A.2 PROOF OF PROPOSITION 2

For fixed i, the softmax denominator is common to all k; exponentiation and division by $\tau > 0$ preserve their ordering. Thus

$$\arg\max_k p_{ik} = \arg\max_k \ell_{ik} = \arg\max_k \widetilde{z}_i^\top \widetilde{c}_k.$$

Because $\|\widetilde{z}_i\|_2 = \|\widetilde{c}_k\|_2 = 1$,

$$\|\widetilde{z}_i - \widetilde{c}_k\|_2^2 = \|\widetilde{z}_i\|_2^2 + \|\widetilde{c}_k\|_2^2 - 2\widetilde{z}_i^\top \widetilde{c}_k = 2 - 2\widetilde{z}_i^\top \widetilde{c}_k.$$

Maximizing the logit therefore minimizes squared Euclidean distance on the unit sphere. Temperature τ changes soft-assignment concentration, but not the hard assignment. □

A.3 PROOF OF THEOREM 1

Step 1: bound each group-summary perturbation. Fix i and h. Both groups contain s variables; r_h reference variables are replaced. Since $|x_{ij}| \le M$, their means satisfy

$$|\widehat{\mu}_{ih} - \mu^*_{ih}| \le \frac{1}{s} \left(\sum_{j \in \mathcal{V}_h^* \setminus \widehat{\mathcal{V}}_h} |x_{ij}| + \sum_{j \in \widehat{\mathcal{V}}_h \setminus \mathcal{V}_h^*} |x_{ij}| \right) \le \frac{2Mr_h}{s}.$$

For the second raw moment $m_{2,ih}$, the same argument gives

$$|\widehat{m}_{2,ih} - m^*_{2,ih}| \le \frac{2M^2 r_h}{s}.$$

The mean absolute value changes by at most $2Mr_h/s$. With $v_{ih} = m_{2,ih} - \mu_{ih}^2$ (divisor s) and both means in $[-M, M]$,

$$\begin{aligned} |\widehat{v}_{ih} - v^*_{ih}| &\le |\widehat{m}_{2,ih} - m^*_{2,ih}| + |\widehat{\mu}^2_{ih} - (\mu^*_{ih})^2| \\ &\le \frac{2M^2 r_h}{s} + |\widehat{\mu}_{ih} - \mu^*_{ih}||\widehat{\mu}_{ih} + \mu^*_{ih}| \\ &\le \frac{2M^2 r_h}{s} + \frac{2Mr_h}{s}(2M) = \frac{6M^2 r_h}{s}. \end{aligned}$$

Combining the four bounds yields

$$\begin{aligned} \|\widehat{b}_{ih} - b^*_{ih}\|_2^2 &\le \left(\frac{r_h}{s}\right)^2 \left\{(2M)^2 + (6M^2)^2 + (2M^2)^2 + (2M)^2\right\} \\ &= \left(\frac{r_h}{s}\right)^2 (8M^2 + 40M^4), \end{aligned}$$

and hence

$$\|\widehat{b}_{ih} - b^*_{ih}\|_2 \le C_M \frac{r_h}{s}, \qquad C_M = \sqrt{8M^2 + 40M^4}.$$

For group-specific sizes s_h, replace s by s_h in each bound.

Step 2: propagate the error to the SGTC input. The PCA scores do not depend on the graph partition and are identical in both representations. Since S_b is L_b-Lipschitz,

$$\|S_b(\widehat{b}_{ih}) - S_b(b^*_{ih})\|_2 \le L_b C_M \frac{r_h}{s}.$$

Concatenation under the Euclidean product norm gives

$$\|\widehat{r}_i - r^*_i\|_2^2 \le \alpha^2 L_b^2 C_M^2 \sum_{h=1}^{H} \left(\frac{r_h}{s}\right)^2,$$

so

$$\|\widehat{r}_i - r^*_i\|_2 \le \alpha L_b C_M \delta_{\mathcal{V}}.$$

Step 3: propagate the error through the encoder and normalization. Adding and subtracting $f_{\theta^*}(\widehat{r}_i)$ gives

$$\begin{aligned}
&\|f_{\widehat{\theta}}(\widehat{r}_i) - f_{\theta^*}(r^*_i)\|_2 \\
&\quad \le \|f_{\widehat{\theta}}(\widehat{r}_i) - f_{\theta^*}(\widehat{r}_i)\|_2 + \|f_{\theta^*}(\widehat{r}_i) - f_{\theta^*}(r^*_i)\|_2 \\
&\quad \le \epsilon_\theta + L_f \alpha L_b C_M \delta_{\mathcal{V}}.
\end{aligned}$$

For any nonzero vectors u and v,

$$\begin{aligned}
\left\| \frac{u}{\|u\|_2} - \frac{v}{\|v\|_2} \right\|_2 &\le \frac{\|u - v\|_2}{\|u\|_2} + \|v\|_2 \left| \frac{1}{\|u\|_2} - \frac{1}{\|v\|_2} \right| \\
&\le \frac{2\|u - v\|_2}{\min(\|u\|_2, \|v\|_2)}.
\end{aligned}$$

The encoder-output norm bound a then gives

$$\|\widehat{\widetilde{z}}_i - \widetilde{z}^*_i\|_2 \le \frac{2}{a}\left(\epsilon_\theta + L_f \alpha L_b C_M \delta_{\mathcal{V}}\right) = \epsilon_z.$$

Step 4: preserve the prototype margin. Add and subtract $(\widetilde{z}^*_i)^\top \widehat{\widetilde{c}}_k$ and apply Cauchy–Schwarz:

$$\begin{aligned}
&\left| \widehat{\widetilde{z}}_i^\top \widehat{\widetilde{c}}_k - (\widetilde{z}^*_i)^\top \widetilde{c}^*_k \right| \\
&\quad \le \|\widehat{\widetilde{z}}_i - \widetilde{z}^*_i\|_2 \|\widehat{\widetilde{c}}_k\|_2 + \|\widetilde{z}^*_i\|_2 \|\widehat{\widetilde{c}}_k - \widetilde{c}^*_k\|_2 \\
&\quad \le \epsilon_z + \epsilon_c.
\end{aligned}$$

The reference similarity can fall, and a competing similarity rise, by at most $\epsilon_z + \epsilon_c$ each. The cosine margin therefore falls by at most $2(\epsilon_z + \epsilon_c)$. If $\kappa > 2(\epsilon_z + \epsilon_c)$, every reference prototype remains the closest in cosine similarity. Proposition 2 preserves all hard prototype labels. □

A.4 Proof of Corollary 1

Let U^* and $\widehat{U}$ contain the first H eigenvectors of L^* and $\widehat{L}$. For eigengap $\gamma = \lambda_{H+1}(L^*) - \lambda_H(L^*) > 0$, Davis–Kahan gives an orthogonal alignment Q with

$$\|\widehat{U}Q - U^*\|_F \le c_{\mathrm{DK}} \frac{\|\widehat{L} - L^*\|_2}{\gamma} = c_{\mathrm{DK}} \frac{\|E\|_2}{\gamma}.$$

If this bound is below the spectral-partition stability radius δ_G, balanced spectral clustering returns the same groups: $r_h = 0$ for all h and $\delta_{\mathcal{V}} = 0$. Theorem 1 gives $\epsilon_z = 2\epsilon_\theta / a$, so

$$\kappa > 2(2\epsilon_\theta / a + \epsilon_c)$$

preserves every prototype label. □

A.5 Proof of Corollary 2

Both centroids average unit-norm embeddings, hence have norm at most one; each embedding–centroid distance is at most two. Set $a_{ik} = \widetilde{\widetilde{z}}_i - \widehat{\mu}_k$ and $b_{ik} = \widetilde{z}_i^* - \mu_k^*$. Then

$$\begin{aligned}\left|\|a_{ik}\|_2^2 - \|b_{ik}\|_2^2\right| &= |\|a_{ik}\|_2 - \|b_{ik}\|_2|(\|a_{ik}\|_2 + \|b_{ik}\|_2) \\ &\leq \|a_{ik} - b_{ik}\|_2(\|a_{ik}\|_2 + \|b_{ik}\|_2) \\ &\leq 4(\epsilon_z + \epsilon_\mu).\end{aligned}$$

The squared reference distance can rise, and a competing distance fall, by at most $4(\epsilon_z + \epsilon_\mu)$ each. Thus the distance margin falls by at most $8(\epsilon_z + \epsilon_\mu)$. If $\Delta > 8(\epsilon_z + \epsilon_\mu)$, every margin stays positive and all final K-means labels are unchanged. □

B Implementation and experimental settings

The six methods were evaluated on the same dataset for each configuration and seed (100–149), giving 50 paired replicates. Table 3 lists the representation and optimization settings.

Table 3: Implementation settings. SGTC-A and SGTC-Σ share the listed optimizer and representation settings; their graphs use thresholded adjacency and precision magnitudes, respectively.

Setting	TC	SGTC-A / SGTC-Σ	UMAP
Input	Standardized X	30 PCA scores and graph summaries	Standardized X
Graph summaries	–	Mean, variance, mean square, mean absolute value	–
Summary weight α	–	0.10	–
Graph groups	–	Eigengap	–
Heads	8	H	–
Hidden layers	[128, 64]	[128, 64]	–
Embedding dimension	64	64	20
Activation / dropout	GELU / 0.1	GELU / 0.1	–
Objective	Prototype CE	Prototype CE	–
Temperature τ	0.5	0.5	–
Pseudo-label initialization	K-means on standardized X	K-means on combined representation	–
Label refresh	Every 15 epochs	Every 15 epochs	–
Optimizer	AdamW	AdamW	–
Learning rate	3×10^{-4}	10^{-4}	–
Weight decay	10^{-4}	10^{-4}	–
Batch size	64	64	–
Maximum epochs	40	80	–
Stopping patience	12	12	–
Neighbors / min. distance	–	–	50 / 0
Final clustering	K-means on normalized embeddings	K-means on normalized embeddings	K-means on coordinates

C ADDITIONAL SIMULATION RESULTS

The sample-size/dimension and dependence/sparsity grids share $(n, m/n, \rho, s) = (500, 5, 0.5, 0.2)$, leaving 17 distinct configurations. The six methods and 50 paired replicates yield 5,100 evaluations of ARI, NMI, FMI, and Purity.

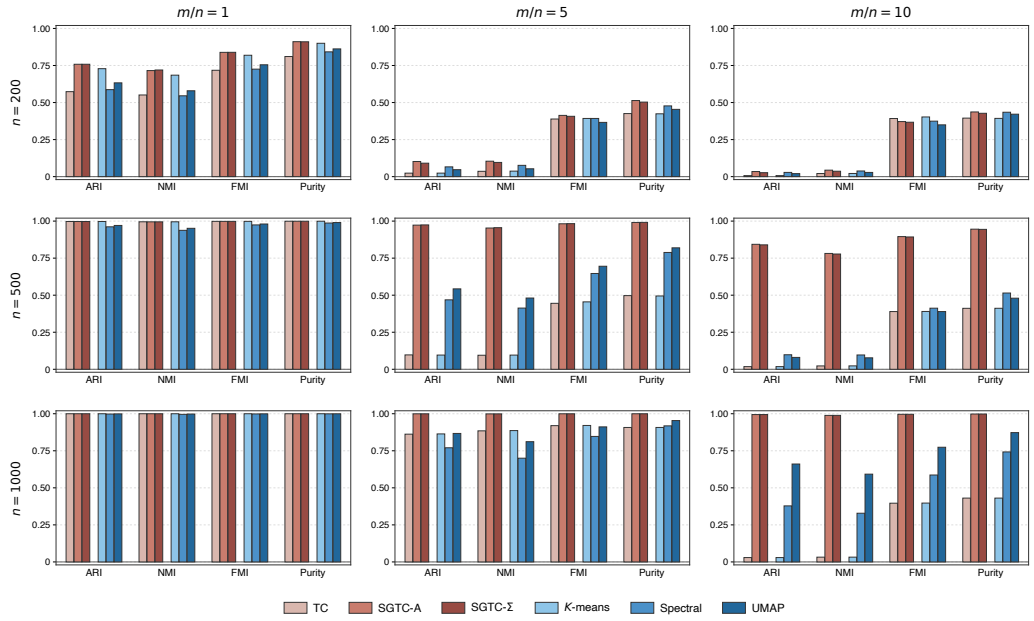

Figure 5: Sample-size/dimension grid: $n \in \{200, 500, 1000\}$, $m/n \in \{1, 5, 10\}$, $\rho = 0.5$, and $s = 0.2$. ARI, NMI, FMI, and Purity summarize 50 paired replicates per setting.

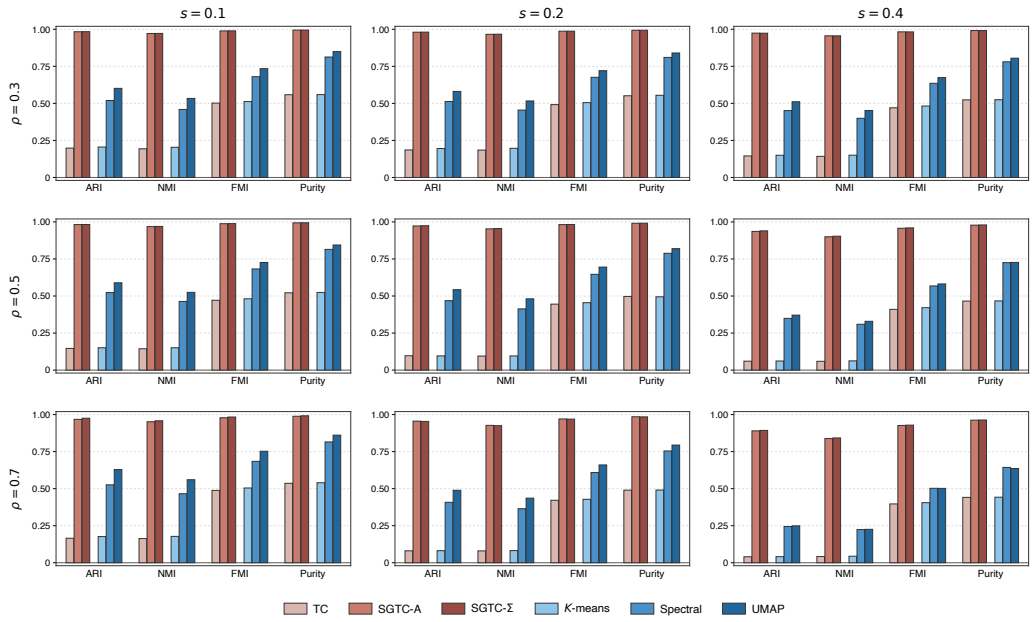

Figure 6: Dependence/sparsity grid: $\rho \in \{0.3, 0.5, 0.7\}$, $s \in \{0.1, 0.2, 0.4\}$, $n = 500$, and $m/n = 5$. ARI, NMI, FMI, and Purity summarize 50 paired replicates per setting.

D GRAPH-PERTURBATION DETAILS

D.1 GRAPH-PERTURBATION RESULTS

At each dimension ratio, both SGTC variants used the same 50 datasets under the unperturbed graph and four matched perturbations. The tables give mean (SD) agreement and mean paired ARI change from the unperturbed fit.

Table 4: Graph perturbations at $n = 500$, $m/n = 1$, $\rho = 0.5$, and $s = 0.2$ (50 paired replicates).

Method	Perturbation	ARI	NMI	FMI	Purity	Mean ΔARI
SGTC-A	Unperturbed	0.997 (0.005)	0.995 (0.009)	0.998 (0.003)	0.999 (0.002)	0.000
SGTC-A	Remove 10%	0.997 (0.005)	0.995 (0.009)	0.998 (0.003)	0.999 (0.002)	-0.000
SGTC-A	Remove 20%	0.997 (0.005)	0.995 (0.009)	0.998 (0.003)	0.999 (0.002)	0.000
SGTC-A	Add 10%	0.997 (0.005)	0.995 (0.008)	0.998 (0.003)	0.999 (0.002)	0.000
SGTC-A	Add 20%	0.997 (0.005)	0.995 (0.008)	0.998 (0.003)	0.999 (0.002)	0.000
SGTC-Σ	Unperturbed	0.997 (0.005)	0.995 (0.009)	0.998 (0.003)	0.999 (0.002)	0.000
SGTC-Σ	Remove 10%	0.997 (0.004)	0.995 (0.008)	0.998 (0.003)	0.999 (0.001)	-0.000
SGTC-Σ	Remove 20%	0.997 (0.005)	0.995 (0.009)	0.998 (0.003)	0.999 (0.002)	-0.000
SGTC-Σ	Add 10%	0.992 (0.028)	0.990 (0.031)	0.995 (0.019)	0.997 (0.010)	-0.005
SGTC-Σ	Add 20%	0.997 (0.005)	0.995 (0.009)	0.998 (0.003)	0.999 (0.002)	0.000

Table 5: Graph perturbations at $n = 500$, $m/n = 5$, $\rho = 0.5$, and $s = 0.2$ (50 paired replicates).

Method	Perturbation	ARI	NMI	FMI	Purity	Mean ΔARI
SGTC-A	Unperturbed	0.973 (0.014)	0.953 (0.024)	0.982 (0.010)	0.991 (0.005)	0.000
SGTC-A	Remove 10%	0.973 (0.014)	0.954 (0.023)	0.982 (0.009)	0.991 (0.005)	0.001
SGTC-A	Remove 20%	0.971 (0.014)	0.951 (0.023)	0.981 (0.010)	0.990 (0.005)	-0.002
SGTC-A	Add 10%	0.973 (0.015)	0.953 (0.024)	0.982 (0.010)	0.991 (0.005)	-0.000
SGTC-A	Add 20%	0.972 (0.014)	0.953 (0.023)	0.982 (0.009)	0.991 (0.005)	-0.000
SGTC-Σ	Unperturbed	0.974 (0.014)	0.956 (0.022)	0.983 (0.009)	0.991 (0.005)	0.000
SGTC-Σ	Remove 10%	0.972 (0.014)	0.953 (0.023)	0.982 (0.009)	0.991 (0.005)	-0.002
SGTC-Σ	Remove 20%	0.968 (0.027)	0.948 (0.035)	0.978 (0.018)	0.989 (0.010)	-0.006
SGTC-Σ	Add 10%	0.972 (0.015)	0.952 (0.024)	0.981 (0.010)	0.991 (0.005)	-0.002
SGTC-Σ	Add 20%	0.972 (0.015)	0.953 (0.024)	0.982 (0.010)	0.991 (0.005)	-0.002

Table 6: Graph perturbations at $n = 500$, $m/n = 10$, $\rho = 0.5$, and $s = 0.2$ (50 paired replicates).

Method	Perturbation	ARI	NMI	FMI	Purity	Mean ΔARI
SGTC-A	Unperturbed	0.843 (0.032)	0.782 (0.035)	0.895 (0.022)	0.946 (0.012)	0.000
SGTC-A	Remove 10%	0.842 (0.035)	0.780 (0.039)	0.895 (0.023)	0.945 (0.013)	-0.001
SGTC-A	Remove 20%	0.838 (0.037)	0.775 (0.040)	0.892 (0.024)	0.943 (0.013)	-0.006
SGTC-A	Add 10%	0.843 (0.036)	0.782 (0.040)	0.895 (0.024)	0.946 (0.013)	0.000
SGTC-A	Add 20%	0.840 (0.034)	0.778 (0.037)	0.893 (0.023)	0.944 (0.013)	-0.003
SGTC-Σ	Unperturbed	0.840 (0.035)	0.778 (0.038)	0.893 (0.024)	0.944 (0.013)	0.000
SGTC-Σ	Remove 10%	0.840 (0.041)	0.778 (0.043)	0.893 (0.027)	0.944 (0.016)	-0.000
SGTC-Σ	Remove 20%	0.844 (0.031)	0.783 (0.035)	0.896 (0.021)	0.946 (0.011)	0.005
SGTC-Σ	Add 10%	0.842 (0.031)	0.780 (0.035)	0.895 (0.021)	0.945 (0.011)	0.003
SGTC-Σ	Add 20%	0.840 (0.038)	0.778 (0.042)	0.893 (0.025)	0.944 (0.014)	-0.000

E COMPUTATIONAL COST

Supplementary Table 7 gives method-call times for SGTC-A and SGTC-Σ across the sample-size and dimension grid. These controlled local measurements are separate from the archived simulation timing records.

Computing environment. We used an Apple M5 MacBook Air with 16 GB of unified memory and one compute thread. Both batches used the same machine, software, thread settings, and SGTC implementation.

Timing protocol. We evaluated $n \in \{200, 500, 1000\}$ and $m/n \in \{1, 5, 10\}$ at $\rho = 0.5$ and $s = 0.2$. Each variant and configuration had one untimed warm-up. Formal runs used seeds 100–109 for $n = 200$ and $n = 1000$ (10 runs each), and 100–119 for $n = 500$ (20 runs). Timing covered PCA, graph-summary construction, neural fitting, cluster assignment, and ARI, NMI, FMI, and Purity calculation; it excluded data generation and shared initial standardization. Feature and graph summaries were recomputed without cross-run cache reuse. We retained all completed formal runs and report sample mean (SD) and median [IQR] in seconds.

Results. Across dimension settings, mean times were 1.3–2.7 seconds at $n = 200$, 3.5–5.2 at $n = 500$, and 13.1–14.6 at $n = 1000$. At $(n, m) = (1000, 10000)$, mean (SD) times were 14.4 (2.9) seconds for SGTC-A and 13.5 (5.1) for SGTC-Σ. Times did not increase monotonically with m at fixed n. This single-machine benchmark of the two SGTC variants does not establish their speed relative to the other clustering methods.

Table 7: SGTC method-call time (s). Each setting had one untimed warm-up; $n = 200$ and $n = 1000$ had 10 formal runs each, and $n = 500$ had 20. Mean (SD) includes all completed runs.

n	m	Method	Runs	Mean (SD)	Median [IQR]
200	200	SGTC-A	50	1.3 (0.7)	1.0 [1.0-1.4]
200	200	SGTC-Σ	50	1.3 (0.5)	1.0 [1.0-1.5]
200	1000	SGTC-A	50	2.2 (1.0)	2.8 [1.1-2.9]
200	1000	SGTC-Σ	50	1.9 (0.9)	2.0 [1.1-2.7]
200	2000	SGTC-A	50	2.2 (1.1)	2.1 [1.3-3.2]
200	2000	SGTC-Σ	50	2.7 (1.1)	3.0 [1.7-3.6]
500	500	SGTC-A	50	4.7 (2.7)	4.1 [3.0-4.9]
500	500	SGTC-Σ	50	5.2 (2.6)	4.2 [3.6-6.3]
500	2500	SGTC-A	50	4.7 (2.1)	4.1 [3.3-5.8]
500	2500	SGTC-Σ	50	4.8 (2.3)	3.5 [2.9-6.3]
500	5000	SGTC-A	50	3.8 (1.8)	3.3 [2.6-4.1]
500	5000	SGTC-Σ	50	3.5 (1.1)	3.3 [3.0-4.0]
1000	1000	SGTC-A	50	13.3 (2.9)	12.7 [10.8-14.8]
1000	1000	SGTC-Σ	50	14.6 (3.5)	15.4 [11.2-17.6]
1000	5000	SGTC-A	50	13.8 (3.0)	15.2 [13.3-15.4]
1000	5000	SGTC-Σ	50	13.1 (3.3)	14.8 [10.7-15.4]
1000	10000	SGTC-A	50	14.4 (2.9)	14.5 [12.9-15.5]
1000	10000	SGTC-Σ	50	13.5 (5.1)	14.9 [9.3-17.0]