Graph Memory: A Structured and Interpretable Framework for Modality-Agnostic Embedding-Based Inference

Artur A. Oliveira[a], Mateus Espadoto[b], Roberto M. Cesar Jr.[c] and Roberto Hirata Jr.[d]
Institute of Mathematics and Statistics, University of São Paulo, Rua do Matão, 1010, São Paulo, Brazil

Keywords: Graph Memory, Multimodal Learning, Reliability Modeling, Interpretability, Histopathology and Gene-Expression Profiles.

Abstract: We introduce **Graph Memory (GM)**, a structured non-parametric framework that represents an embedding space as a compact graph of reliability-annotated prototype regions. GM encodes local geometry and regional ambiguity through prototype relations and performs inference by diffusing query evidence over this graph, unifying instance retrieval, prototype-based reasoning, and graph diffusion within a single inductive and interpretable model. The framework is modality-agnostic: in multimodal settings, independent prototype graphs are built per modality and combined through reliability-aware late fusion. Experiments on synthetic benchmarks, breast histopathology (IDC), and the multimodal AURORA dataset show that GM matches or exceeds *k*NN and Label Spreading in accuracy, while delivering substantially better calibration, smoother decision boundaries, and an order-of-magnitude smaller memory footprint. Overall, GM provides a principled and interpretable approach to non-parametric inference across single- and multi-modal domains.

1 INTRODUCTION

Embedding-based classifiers are commonly built around two paradigms: *parametric* heads that learn global decision boundaries, and *non-parametric* methods such as prototypes or *k*NN that rely on local geometric similarity. While effective, both paradigms largely treat the embedding space as a collection of independent points or loosely defined clusters, with limited ability to model relations between regions, capture where uncertainty concentrates, or propagate evidence across decision boundaries. As a result, broader relational structure and multi-hop dependencies in embedding spaces are often ignored.

We introduce **Graph Memory (GM)**, a compact and reusable representation that explicitly organizes an embedding space into a graph of region-level prototypes. Each prototype summarizes a coherent region through its centroid and reliability attributes (e.g., purity, dispersion, stability), while edges encode geometric proximity and persistent ambiguity between regions. This reliability-weighted prototype graph provides an interpretable abstraction of the embedding manifold and supports context-aware inference that integrates local similarity with global structure.

Inference in GM proceeds by activating prototypes relevant to a query and diffusing this activation over the prototype graph, yielding calibrated and geometrically smooth predictions. The diffusion process preserves interpretability: prototypes accumulating the most evidence directly expose which regions support or challenge a decision. In this way, GM unifies instance retrieval, prototype-based reasoning, and graph diffusion within a single inductive framework, without relying on full instance graphs or transductive inference.

A key strength of GM is its effectiveness in heterogeneous, sparsely sampled, or high-dimensional regimes, where classical non-parametric methods struggle to capture global structure without excessive memory or computation. This naturally extends to multimodal settings: GM constructs independent prototype graphs per modality and combines their predictions through reliability-aware late fusion, preserving modality-specific interpretability while enabling principled integration of complementary evidence.

We evaluate GM across increasingly challenging scenarios. On controlled synthetic benchmarks, GM

[a] https://orcid.org/0000-0002-3606-1687
[b] https://orcid.org/0000-0002-1922-4309
[c] https://orcid.org/0000-0003-2701-4288
[d] https://orcid.org/0000-0003-3861-7260

Oliveira, A. A., Espadoto, M., Cesar Jr., R. M. and Hirata Jr., R.
Graph Memory: A Structured and Interpretable Framework for Modality-Agnostic Embedding-Based Inference.
DOI: 10.5220/0014578800004084

In *Proceedings of the 21st International Conference on Computer Vision Theory and Applications (VISAPP 2026) - Volume 1*, pages 652-659
ISBN: 978-989-758-804-4; ISSN: 2184-4321

yields smoother and better-calibrated decision functions than classical *k*NN and Label Spreading (Pedregosa et al., 2011). On high-dimensional breast histopathology (IDC) data (Janowczyk and Madabhushi, 2016; Cruz-Roa et al., 2014), GM achieves strong accuracy and competitive calibration using a compact prototype set. Finally, on the multimodal AURORA dataset (Garcia-Recio et al., 2023), combining whole-slide images and RNA-seq profiles, GM improves calibration through reliability-aware fusion and remains robust in an extremely small-sample regime.

Overall, Graph Memory provides a compact, interpretable, and modality-agnostic framework for embedding-based inference, supporting smooth decision boundaries, calibrated predictions, and principled multimodal integration within a unified inductive model.

2 RELATED WORK

Prior work on embedding-space inference spans parametric classifiers, non-parametric retrieval, prototype-based models, and graph-based semi-supervised learning. Parametric heads learn global decision boundaries, while non-parametric approaches such as prototypes and *k*NN rely on local neighborhood evidence. Despite their effectiveness, both families typically lack an explicit, reusable representation of how *regions* of the embedding space relate, where ambiguity concentrates, and how evidence should propagate across regions.

Prototype-based classifiers summarize classes via representative centroids or parts, as in nearest-class-mean methods, Prototypical Networks (Snell et al., 2017), and ProtoPNet-style interpretability models (Chen et al., 2019). These approaches improve interpretability and data efficiency, but generally treat prototypes in isolation, without modeling persistent relations, cross-class ambiguity, or reliability differences between regions.

Instance-based retrieval methods range from classical *k*NN to Deep *k*NN (Papernot and McDaniel, 2018) and more recent semi-parametric adapters that integrate retrieval with pretrained encoders for adaptation, privacy, or domain transfer (Doerrich et al., 2024; Bhardwaj et al., 2023; Huang et al., 2023). While effective, these methods operate primarily at the instance or activation level and are tightly coupled to specific architectures, limiting reuse as a general inference layer.

Graph-based semi-supervised learning (Graph SSL) propagates labels over instance graphs (Zhu et al., 2003), with Graph Convolutional Networks extending this idea via learned message passing (Kipf and Welling, 2017). Diffusion-based methods such as Personalized PageRank (Haveliwala, 2002) highlight the benefits of restart-based propagation for personalized inference. However, most graph-based approaches rely on instance-level graphs, are transductive, and leave graph construction and inductive query attachment under-specified (Zhou et al., 2003).

Beyond pairwise distances, richer relations have been explored through temporal graphs (Chanpuriya et al., 2022), and scalable non-parametric retrieval systems such as MUVERA (Dhulipala et al., 2024). These works demonstrate expressive relations and efficiency at scale, but do not provide a compact, region-level memory that jointly encodes geometry, reliability, and ambiguity for inductive classification.

Interpretability and reliability have been studied through prototype-based explanations (Chen et al., 2019) and post-hoc calibration techniques, including temperature scaling (Guo et al., 2017), Dirichlet calibration (Kull et al., 2019), and energy-based OOD detection (Liu et al., 2020). Orthogonally, clustering validity indices such as Silhouette, Davies–Bouldin, Dunn, and Calinski–Harabasz characterize complementary aspects of region quality (Rousseeuw, 1987; Davies and Bouldin, 1979; Dunn, 1974; Caliński and Harabasz, 1974; Liu et al., 2010).

Summary of Gaps. Taken in isolation, existing approaches typically (i) operate at the instance or prototype level without modeling inter-region relations, (ii) lack a persistent and reusable region-level memory, (iii) omit reliability and ambiguity as first-class signals, or (iv) depend on transductive graphs or architecture-specific adapters. These limitations motivate a structured, prototype-level graph memory that explicitly encodes regional geometry, reliability, and relations, while remaining inductive, interpretable, and compatible with modern embedding-based inference. These gaps motivate a representation that explicitly models regions, their relations, and their reliability in a reusable and inductive manner.

3 METHOD

Graph Memory (GM) is a reusable, non-parametric representation that models an embedding space as a graph of prototype regions. Inference is performed via *evidence diffusion*: a query induces an initial activation over prototypes, which is propagated through a reliability-weighted random walk with restart. The resulting stationary distribution integrates local sim-

ilarity with global structure while preserving query-dependent inference.

3.1 Overview

Given a labeled embedding dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i = f(x_i^{\text{raw}}) \in \mathbb{R}^d$ are fixed embeddings and $y_i \in \{1, \ldots, C\}$ are class labels, we construct a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ whose nodes $\mathcal{V} = \{1, \ldots, K\}$ are cluster-level prototypes. Each prototype summarizes a subset $S_c \subseteq \mathcal{D}$ and edges $(c, c') \in \mathcal{E}$ encode relational affinity. The pipeline has three parts: (1) joint prototype construction, (2) relation graph formation, and (3) inductive inference by graph diffusion.

3.2 Prototype Selection

We partition the embeddings $\{x_i\}$ into K groups using K-means or K-medoids across all classes jointly, preserving mixed boundary regions that per-class clustering would treat separately. Each prototype c is represented by its centroid

$$\mu_c = \frac{1}{|S_c|} \sum_{x_i \in S_c} x_i,$$

and the following attributes are stored:

- **Support:** $|S_c|$, number of samples represented by prototype c.
- **Dominant class:** $y_c = \arg\max_y |\{x_i \in S_c : y_i = y\}|$, the most frequent class within S_c.
- **Purity:** $\pi_c = \frac{\max_y |\{x_i \in S_c : y_i = y\}|}{|S_c|}$, the proportion of samples belonging to the dominant class, used later as a reliability factor.

All statistics are normalized by the local population $|S_c|$, ensuring that prototype attributes reflect the stability of the region they summarize.

3.3 Prototype Quality and Reliability

Each prototype is further characterized by geometric and stability metrics that are combined into a scalar reliability score $r_c \in [0, 1]$.

(1) Normalized silhouette. For each sample $x_i \in S_c$, let

$$a(i) = \frac{1}{|S_c|-1} \sum_{x_j \in S_c,\, j \neq i} \|x_i - x_j\|_2, \qquad \text{(intra-cluster dist.)} \tag{1}$$

$$b(i) = \min_{c' \neq c} \frac{1}{|S_{c'}|} \sum_{x_k \in S_{c'}} \|x_i - x_k\|_2, \qquad \text{(nearest-cluster dist.)} \tag{2}$$

and compute the normalized silhouette value

$$s_c = \frac{1}{|S_c|} \sum_{x_i \in S_c} \frac{\text{sil}(i) + 1}{2}, \quad \text{where sil}(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}.$$

This rescales the original silhouette from $[-1, 1]$ to $[0, 1]$, aligning its range with the other metrics.

(2) Dispersion. The internal variance relative to the centroid:

$$v_c = \frac{1}{|S_c|} \sum_{x_i \in S_c} \|x_i - \mu_c\|_2^2.$$

Low v_c indicates compact geometry, while high values imply intra-class spread or outliers.

(3) Margin. Separation from prototypes of different dominant classes:

$$m_c = \min_{c' : y_{c'} \neq y_c} \|\mu_c - \mu_{c'}\|_2.$$

(4) Instability. Sensitivity to embedding perturbations. We sample $x_i^{\delta} = x_i + \delta_i$ (small Gaussian noise) and measure reassignment rate:

$$\rho_c = \frac{1}{|S_c|} \sum_{x_i \in S_c} \mathbf{1}\left[\text{NN}(x_i^{\delta}) \neq c\right].$$

Lower ρ_c means more stable assignments.

(5) Robust normalization. Among the prototype metrics, only the margin m_c and dispersion v_c are unbounded and directly depend on the scale of the embedding space. To make them comparable with the bounded quantities (s_c, ρ_c, π_c), we apply a robust rescaling based on median and interquartile range:

$$\bar{x}_c = \sigma\left(\frac{x_c - \text{med}(x)}{\text{IQR}(x)}\right) \tag{3}$$

$$\text{IQR}(x) = Q_{0.75}(x) - Q_{0.25}(x)$$

where $\sigma(\cdot)$ is the logistic function. This transformation, akin to robust scaling (Rousseeuw and Croux, 1993), compresses extreme values and removes sensitivity to global scale, ensuring that large distances or variances do not dominate the composite reliability. For metrics already bounded in $[0, 1]$ (e.g., s_c, ρ_c, π_c), no additional normalization is applied.

(6) Composite reliability. The overall reliability is then

$$r_c = \sigma\Big(\lambda_1 s_c + \lambda_2 \bar{m}_c + \lambda_5 \pi_c - \lambda_3 \rho_c - \lambda_4 \bar{v}_c\Big),$$

where λ_i are non-negative weights (default $= 1$). Bounded metrics (s_c, ρ_c, π_c) are used directly, while unbounded metrics (m_c, v_c) are robustly normalized as in Eq. 3. This aggregation balances compactness, separation, stability, and purity on a comparable numerical scale.

3.4 Graph Memory Construction

Prototype relations are encoded in a weighted graph built in the embedding space, where edges reflect both geometric affinity and contextual reliability between regions. Each edge weight measures the local affinity between prototypes c and c' through a Gaussian kernel:

$$A_{cc'} = \begin{cases} \exp(-\beta \|\mu_c - \mu_{c'}\|_2^2), & c' \in \mathrm{KNN}_k(c), \\ 0, & \text{otherwise.} \end{cases}$$

The adjacency matrix A is symmetrized and row-normalized to obtain a stochastic transition matrix

$$S = D^{-1}A, \qquad D_{cc} = \sum_{c'} A_{cc'},$$

which defines the propagation operator used during inference.

3.5 Inference by Graph Diffusion

Given a query embedding x_q, GM first computes an initial activation over prototypes,

$$z_0(c) = \exp(-\beta \|x_q - \mu_c\|_2^2)\, r_c,$$

where μ_c is the prototype centroid and r_c its reliability. This z_0 serves as a *starting distribution*, indicating how strongly the query is associated with each prototype.

Instead of predicting directly from these local affinities, GM performs *evidence diffusion* on the prototype graph. The activation is propagated by a random-walk–with-restart update,

$$z^{(t+1)} = (1-\alpha)\, z_0 + \alpha S z^{(t)},$$

where S is the row-normalized affinity matrix and $\alpha \in [0,1)$ controls how much evidence flows through the graph at each step. With probability α the walk transitions according to S, and with probability $1-\alpha$ it restarts from z_0. This process is guaranteed to converge to a unique stationary activation vector

$$z = (I - \alpha S)^{-1} z_0, \tag{4}$$

which corresponds to the steady state of the classical label-propagation formulation of Zhou *et al.* (Zhou et al., 2003).

Class probabilities are obtained by aggregating stationary activations over prototypes with the same dominant label:

$$p(y \mid x_q) \propto \sum_{c:\, y_c = y} z_c, \qquad \hat{y}_q = \arg\max_y p(y \mid x_q).$$

When $\alpha = 0$, GM reduces to a purely local scheme analogous to weighted kNN. As α increases, diffusion incorporates more relational information from the prototype graph, yielding smoother decision boundaries and more calibrated predictions.

3.6 Multimodal Inference with Independent Graph Memories

Graph Memory (GM) is inherently modality-agnostic and operates on embeddings from any fixed encoder. In multimodal settings, where heterogeneous feature spaces (e.g., images, genomic profiles, clinical data) make early fusion brittle due to differences in scale, noise, and sampling density, GM constructs **independent graph memories** for each modality. Each memory performs prototype-level diffusion to produce calibrated posteriors, which are combined through reliability-aware late fusion, integrating complementary evidence without cross-modal alignment while preserving modality-level interpretability.

Independent Memories. Let $\mathcal{X} = \{x^{(1)}, \ldots, x^{(M)}\}$ denote M aligned modalities describing the same set of N samples, and let $f_m(\cdot)$ be the fixed encoder for modality m. Each encoder produces embeddings $h_i^{(m)} = f_m(x_i^{(m)}) \in \mathbb{R}^{d_m}$. For every modality, we construct an independent graph memory

$$\mathrm{GM}_m = (\mathcal{P}_m, S_m, r_m),$$

where $\mathcal{P}_m = \{\mu_c^{(m)}\}_{c=1}^{K_m}$ are prototype centroids, S_m is the row-stochastic diffusion matrix derived from pairwise affinities among prototypes, and r_m are reliability weights computed as described in Sec. 3. Each GM_m operates on its own embedding space, encoding how local regions relate and how reliable they are. Inference for a query sample $x_q^{(m)}$ proceeds independently within each modality:

$$p_m(y \mid x_q^{(m)}) \propto \sum_{c:\, y_c^{(m)} = y} z_c^{(m)}, \; z^{(m)} = (I - \alpha S_m)^{-1} z_0^{(m)},$$

where the initial activation vector is

$$z_0^{(m)}(c) = \exp\big(-\beta_m \|h_q^{(m)} - \mu_c^{(m)}\|_2^2\big)\, r_c^{(m)}.$$

Thus, each modality yields an independent, reliability-weighted posterior $p_m(y)$ that reflects its own geometric structure and evidence distribution.

Interpretable Consensus. Rather than enforcing early fusion in the embedding space, GM retains separate posteriors $\{p_m(y)\}_{m=1}^{M}$ that can be compared or combined at the decision level. A reliability-aware consensus prediction is obtained as

$$p_{\text{fused}}(y) = \frac{\sum_m \omega_m r_m^{(q)}\, p_m(y)}{\sum_m \omega_m r_m^{(q)}}, \tag{5}$$

where $r_m^{(q)}$ denotes the mean reliability of prototypes activated for query q in modality m, and ω_m is an

optional scalar *modality weight* controlling the relative influence of each modality. If no prior preference is given, ω_m=1 for all m, yielding an unweighted reliability-based fusion. This late-fusion formulation preserves interpretability by keeping per-modality reasoning explicit, while still enabling a unified consensus prediction when desired.

Cross-Modal Agreement. Disagreement between modalities is informative rather than undesirable. For any pair of modalities (A,B), we define a simple consistency index

$$A_{AB} = 1 - \tfrac{1}{2}\sum_{y} |p_A(y) - p_B(y)|,$$

where low A_{AB} indicates high disagreement and potential uncertainty. Tracking agreement across modalities exposes complementary or conflicting evidence, enabling analyses of uncertainty, redundancy, and complementary signal across feature spaces.

4 EXPERIMENTAL EVALUATION

The experiments evaluate Graph Memory's (GM) accuracy, calibration, smoothness, and interpretability, as well as its robustness to compression, imbalance, and multimodal heterogeneity. We evaluate GM on synthetic benchmarks, breast histopathology (IDC), and multimodal biomedical data (AURORA), focusing on accuracy, calibration, smoothness, and interpretability. All methods operate on the same frozen embeddings, train/test splits, and preprocessing. Hyperparameters are selected on validation data. Our experimental code is available at https://github.com/arturandre/clip-qr-kg

Metrics. We report Top-1 accuracy, negative log-likelihood (NLL), and the smoothness of decision functions via Dirichlet energy. All results are reported as 'mean (std)' over ten runs. For 2D datasets, smoothness is estimated from the mean squared gradient of the class-probability field:

$$E_{\mathrm{2D}} = \mathbb{E}_{(x,y)\in\mathcal{G}}\left[\|\nabla p(y{=}1 \mid x)\|_2^2\right],$$

where lower values indicate smoother and more stable decision transitions. For higher-dimensional embeddings, the same principle is measured via the *graph Dirichlet energy*, a standard notion of smoothness in signal processing on graphs (Shuman et al., 2013):

$$E(f) = \tfrac{1}{2|\mathcal{E}|}\sum_{(i,j)\in\mathcal{E}} w_{ij}\left(f_i - f_j\right)^2$$

$$w_{ij} = \exp(-\beta\,\|x_i - x_j\|_2^2)$$

where $f_i = p(y{=}1 \mid x_i)$ and $\mathcal{E}$ denotes k-nearest-neighbor edges. Lower energy values correspond to smoother, more regular decision functions across the embedding manifold.

4.1 Baselines

We compare GM to representative parametric and non-parametric models: **Linear probe:** multinomial logistic regression, ***k*NN:** instance-level k-nearest neighbors, **Budget-*k*NN:** kNN trained on a stratified subset whose size matches GM's prototype budget ($|\text{train}|{=}K$), and **Label Spreading (LS):** scikit-learn transductive label spreading on an instance graph with n-neighbors (Pedregosa et al., 2011).

4.2 Synthetic Datasets

We use the *moons* and *circles* datasets (Pedregosa et al., 2011) (4000 samples each, 50/50 split). Class-imbalance experiments downsample the minority class to an 8:1 ratio. GM uses $K = 120$ prototypes (balanced) or $K = 112$ (imbalanced) and k_{graph}=10, α=0.5, β=0.1, and attach-k=8. Figure 1 shows decision regions and gradient maps.

Table 1 shows that GM matches kNN and LS in accuracy but achieves consistently lower NLL, using only $\sim 10\%$ of training samples. GM remains fully *inductive*, unlike LS, and robust under long-tail imbalance. Budget-kNN highlights that naive compression severely hurts performance, whereas GM remains stable due to structured region prototypes. Table 2 reports smoothness: GM reduces Dirichlet energy by $30 \sim 40$**%** relative to kNN and by **94%** relative to LS, confirming flatter, more reliable boundaries.

4.3 Breast Histopathology (IDC)

We evaluate Graph Memory (GM) on the *Breast Histopathology Images (Invasive Ductal Carcinoma, IDC)* dataset (binary: IDC vs. non-IDC)[1] (Janowczyk and Madabhushi, 2016; Cruz-Roa et al., 2014). A ResNet18 encoder provides frozen 512-D embeddings. GM uses K=32 prototypes with the same hyperparameters as in Sec. 4.2. Table 3 shows that GM attains the highest accuracy and near-parametric NLL despite using only 32 prototypes. Table 4 shows that GM also achieves the lowest graph Dirichlet energy, indicating stable, smooth decision functions.

[1] Available at https://huggingface.co/datasets/dbzadnen/breast-histopathology-images

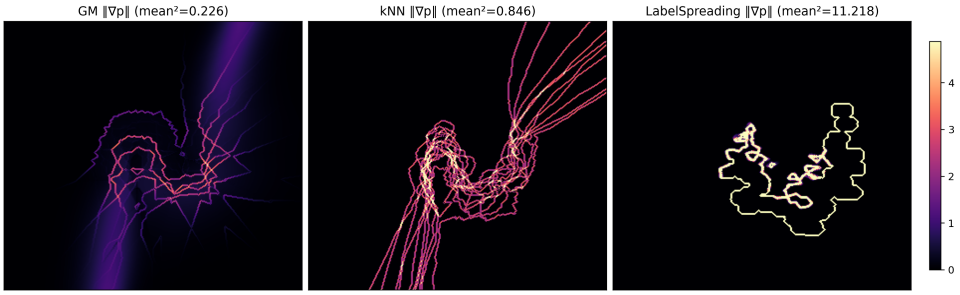

Figure 1: **Dirichlet energy maps** (visualized as gradient-magnitude fields, $\|\nabla p(y{=}1 \mid x)\|$). Lower values indicate smoother and more stable decision transitions. GM produces lower-energy, reliability-weighted boundaries and avoids the sample-hugging artifacts observed in Label Spreading and *k*NN, particularly near class interfaces.

Table 1: **Accuracy and calibration on synthetic binary datasets.** GM matches *k*NN and Label Spreading in accuracy while achieving better calibration and requiring only ∼10% of the samples. Unlike *k*NN and Label Spreading, GM remains inductive and robust under the 8:1 long-tail regime.

Dataset	Imbal.?	Method	Acc ↑	NLL ↓
moons	No	GM (P=120) *Ours*	0.936 (0.008)	**0.178 (0.010)**
		kNN-budget(P=120)	0.935 (0.006)	0.298 (0.077)
		kNN	**0.940 (0.007)**	0.402 (0.062)
		Label Spreading	0.936 (0.010)	0.226 (0.018)
		Linear	0.857 (0.005)	0.318 (0.006)
	Yes	GM (P=112) *Ours*	0.871 (0.018)	**0.339 (0.052)**
		kNN-budget(P=112)	0.689 (0.033)	1.153 (0.414)
		kNN	0.880 (0.012)	0.758 (0.192)
		Label Spreading	**0.889 (0.014)**	0.526 (0.065)
		Linear	0.780 (0.011)	0.531 (0.026)
circles	No	GM (P=120) *Ours*	**0.947 (0.005)**	**0.178 (0.009)**
		kNN-budget(P=120)	0.936 (0.007)	0.353 (0.024)
		kNN	**0.947 (0.004)**	0.331 (0.043)
		Label Spreading	0.940 (0.006)	0.206 (0.028)
		Linear	0.499 (0.008)	0.693 (0.000)
	Yes	GM (P=112) *Ours*	0.859 (0.014)	**0.327 (0.013)**
		kNN-budget(P=112)	0.857 (0.015)	0.543 (0.132)
		kNN	**0.892 (0.008)**	0.654 (0.041)
		Label Spreading	0.883 (0.007)	0.472 (0.018)
		Linear	0.500 (0.000)	1.158 (0.000)

4.4 Multimodal Medical Benchmark

We evaluate Graph Memory (GM) on the AURORA program (Garcia-Recio et al., 2023), a longitudinal, multi-institutional cohort profiling paired primary and metastatic breast tumors. Its RNA-seq component (GSE209998)[2] (Garcia-Recio et al., 2023; Zahraeifard et al., 2024; Garcia-Recio et al., 2025; Edwards et al., 2025) provides patient-matched gene-expression data with rich clinical annotation.

We embed whole-slide images with TITAN (Ding et al., 2024). Median aggregation yields a single patient-level descriptor. Gene-expression profiles are filtered to remove extremely low-variance genes, aggregated via median to represent one genomic vector per patient. We use **ResNet18 encoders** (trained for 20 epochs on a fixed 50/50 split), replacing the final layer with a 256-D projection head) to extract same dimensionality embeddings for both modalities. These embeddings serve as the input to the corresponding Graph Memory.

Inference is performed independently on each modality. Final predictions use the **reliability-weighted late fusion** rule from Eq. 5, combining morphological (WSI) and molecular (RNA-seq) evidence without early-fusion assumptions. Table 5

[2]https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE209998

Table 2: **Graph Dirichlet energy comparison (lower is better).** The metric quantifies the average squared gradient magnitude of the class-probability field, where smaller values denote smoother decision transitions. **Graph Memory (GM)** yields substantially lower energy (**30-40%** below *k*NN and over **94%** below Label Spreading) indicating smoother, more reliable boundaries while preserving accuracy and calibration.

Dataset	Imbal.?	Method	E_{2D} ↓
moons	No	GM (P=120)	**0.927 (0.048)**
		kNN	1.410 (0.095)
		Label Spreading	17.705 (1.322)
	Yes	GM (P=112)	**0.834 (0.054)**
		kNN	1.177 (0.066)
		Label Spreading	14.216 (1.275)
circles	No	GM (P=120)	**0.723 (0.063)**
		kNN	1.065 (0.088)
		Label Spreading	11.457 (0.786)
	Yes	GM (P=112)	**0.403 (0.044)**
		kNN	0.663 (0.036)
		Label Spreading	6.500 (0.726)

Table 3: **Accuracy and calibration on the IDC dataset.** GM achieves the highest accuracy and competitive calibration using only 32 prototypes, demonstrating compact and reliable non-parametric inference.

Method	Acc ↑	NLL ↓
GM(P=32) (Ours)	**0.845 (0.000)**	0.420 (0.000)
kNN-budget(P=32)	0.797 (0.078)	0.607 (0.339)
kNN	0.840 (0.000)	1.039 (0.000)
Label Spreading	0.826 (0.000)	0.531 (0.000)
Linear	0.836 (0.000)	**0.397 (0.000)**

Table 4: **Graph Dirichlet energy on the IDC dataset.** Lower values indicate smoother and more stable decision functions. Graph Memory (GM) achieves the lowest energy, yielding smoother decision boundaries than *k*NN and Label Spreading while maintaining negligible variance across runs.

Method	graph Dirichlet energy
GM (P=32) (Ours)	0.291 (0.000)
kNN	0.321 (0.000)
Label Spreading	0.344 (0.000)

summarizes accuracy and NLL. Both modalities perform well individually, fusion consistently improves calibration and accuracy. Table 6 reports the corresponding graph Dirichlet energy, showing that both memories remain smooth and stable despite the extremely small sample size.

Table 5: **AURORA multimodal results.** Late fusion improves accuracy and calibration by integrating complementary morphological and molecular evidence while preserving per-modality interpretability.

Method	Acc ↑	NLL ↓
GM_1 (WSI)	0.933 (0.034)	0.917 (0.561)
GM_2 (Genetic)	0.951 (0.039)	0.236 (0.279)
GM-fused	**0.971 (0.013)**	**0.120 (0.025)**

Table 6: **Graph Dirichlet energy on AURORA.** Both memories exhibit low energy and modest variance across runs, indicating smooth and stable decision functions even in the small-sample regime.

Method	Dirichlet energy ↓
GM_1 (WSI)	0.634 (0.111)
GM_2 (Genetic)	0.514 (0.122)

5 DISCUSSION AND CONCLUSIONS

Graph Memory (GM) addresses the lack of a structured, reusable representation that captures region-level geometry, and relational context combining prototype abstraction, reliability modeling, and graph-based diffusion. GM subsumes several classical non-parametric methods as limiting cases. When each prototype represents a single instance and diffusion is disabled (α=0), GM reduces to standard *k*NN; when each class collapses to a single prototype with uniform reliability ($r_c \equiv 1$), it recovers nearest-centroid classification; and when diffusion is active with uniform reliabilities, GM yields the steady-state solution of a label-propagation system.

Unlike these baselines, GM operates on a compact set of region-level prototypes connected by a reliability-weighted graph that encodes both local geometry (nodes) and contextual relations (edges), enabling calibrated, context-aware, and geometrically smooth predictions while remaining fully inductive and embedding-agnostic. These theoretical properties are reflected empirically: GM consistently matches or outperforms *k*NN, Budget-*k*NN, and Label Spreading across synthetic benchmarks, histopathology (IDC), and multimodal biomedical data (AURORA), while yielding improved calibration and smoother decision boundaries. In contrast to instance-level retrieval and transductive graph SSL methods (Zhu et al., 2003; Kipf and Welling, 2017), GM relies on region summaries rather than full instance graphs, enabling stable inference in low-sample and high-dimensional regimes. The number of prototypes K controls the granularity of this representation, interpolating between instance-level retrieval for large K and coarser region abstraction for smaller K, with stable behavior across a broad range of values when the embedding manifold is adequately covered.

Overall, Graph Memory offers a compact and fully inductive alternative to instance-based retrieval and graph SSL, delivering calibrated predictions, smooth decision boundaries, substantial memory savings over instance-level methods, and principled multimodal integration via reliability-weighted late fusion, without test-time graph reconstruction.

ACKNOWLEDGMENTS

This work was funded partially by FAPESP project 2022/15304-4 and MCTI (law 8.248, PPI-Softex - TIC 13 - 01245.010222/2022-44). The results published here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

REFERENCES

Bhardwaj, H., Pruthi, D., and Neubig, G. (2023). knn-cm: Semi-parametric classifiers via *k*nn-based class-conditional memory. In *Findings of the Association for Computational Linguistics: EMNLP*.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27.

Chanpuriya, S. et al. (2022). Direct embedding of temporal network edges via time-decayed line graphs.

Chen, C. et al. (2019). This looks like that: Deep learning for interpretable image recognition. In *NeurIPS*.

Cruz-Roa, A. et al. (2014). Automatic detection of invasive ductal carcinoma in whole slide images with convo-

lutional neural networks. In *Medical imaging 2014: Digital pathology*, volume 9041, page 904103. SPIE.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.

Dhulipala, L., Hadian, M., Jayaram, R., Lee, J., and Mirrokni, V. (2024). Muvera: Multi-vector retrieval via fixed dimensional encodings.

Ding, T. et al. (2024). Multimodal whole slide foundation model for pathology.

Doerrich, S. et al. (2024). Integrating knn with foundation models for adaptable and privacy-aware image classification. *arXiv preprint arXiv:2402.12500*.

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions.

Edwards, D. N. et al. (2025). Increased fatty acid delivery by tumor endothelium promotes metastatic outgrowth. *JCI Insight*, 10(9):e187531.

Garcia-Recio, S. et al. (2023). Multiomics in primary and metastatic breast tumors from the aurora us network finds microenvironment and epigenetic drivers of metastasis. *Nature Cancer*, 4(1):128–147.

Garcia-Recio, S. et al. (2025). Understanding metastasis mixed-treatment responses through genomic analyses. *NPJ Breast Cancer*, 11(1):9.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *ICML*.

Haveliwala, T. H. (2002). Topic-sensitive pagerank. In *Proceedings of the 11th international conference on World Wide Web*, pages 517–526.

Huang, Y., Liu, D., Zhong, Z., Shi, W., and Lee, Y. T. (2023). knn-adapter: Efficient domain adaptation for black-box language models. In *arXiv preprint arXiv:2302.10879*.

Janowczyk, A. and Madabhushi, A. (2016). Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7(1):29.

Kipf, T. and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

Kull, M. et al. (2019). Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. *NeurIPS*.

Liu, W., Wang, X., Owens, J. D., and Finn, C. (2020). Energy-based out-of-distribution detection. In *NeurIPS*.

Liu, Y. et al. (2010). Understanding of internal clustering validation measures. In *2010 IEEE International Conference on Data Mining*, pages 911–916.

Papernot, N. and McDaniel, P. (2018). Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. In *International Conference on Learning Representations*.

Pedregosa, F. et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rousseeuw, P. and Croux, C. (1993). Alternatives to median absolute deviation. *Journal of the American Statistical Association*, 88:1273 – 1283.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Shuman, D. I. et al. (2013). The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30(3):83–98.

Snell, J., Swersky, K., and Zemel, R. (2017). Prototypical networks for few-shot learning. In *NeurIPS*.

Zahraeifard, S. et al. (2024). Loss of tumor suppressors promotes inflammatory tumor microenvironment and enhances lag3+ t cell mediated immune suppression. *Nature Communications*, 15(1):5873.

Zhou, D., Bousquet, O., Lal, T., Weston, J., and Schölkopf, B. (2003). Learning with local and global consistency. *Advances in neural information processing systems*, 16.

Zhu, X., Ghahramani, Z., and Lafferty, J. (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*.