Knowledge Graphs as Structured Memory for Embedding Spaces: From Training Clusters to Explainable Inference

Artur A. Oliveira[1][a], Mateus Espadoto[1][b], Roberto M. Cesar Jr.[1][c] and Roberto Hirata Jr.[1][d]

[1]*Institute of Mathematics and Statistics, University of São Paulo, Rua do Matão, 1010, São Paulo, Brazil*
{arturao, mespadot, hirata}@ime.usp.br, rmcesar@usp.br

Keywords:
Graph Memory; Prototype Learning; Calibration; Interpretability; Histopathology

Abstract:
We introduce **Graph Memory (GM)**, a structured non-parametric framework that augments embedding-based inference with a compact, relational memory over region-level prototypes. Rather than treating each training instance in isolation, GM summarizes the embedding space into prototype nodes annotated with reliability indicators and connected by edges that encode geometric and contextual relations. This design unifies instance retrieval, prototype-based reasoning, and graph-based label propagation within a single inductive model that supports both efficient inference and faithful explanation. Experiments on synthetic and real datasets including breast histopathology (IDC) show that GM achieves accuracy competitive with kNN and Label Spreading while offering substantially better calibration and smoother decision boundaries, all with an order of magnitude fewer samples. By explicitly modeling reliability and relational structure, GM provides a principled bridge between local evidence and global consistency in non-parametric learning.

1 INTRODUCTION

Modern embedding-based classifiers are typically realized in two complementary paradigms: *parametric* models (e.g., linear-softmax heads) that learn global decision boundaries, and *non-parametric* approaches (e.g., prototypes, kNN) that infer predictions from local neighborhoods in the embedding space. While both are effective in practice, they usually treat the embedding space as a collection of isolated samples or clusters, without explicitly modeling how different regions relate to one another, where uncertainty persists, or how local evidence should interact across boundaries. Simple aggregate measures such as purity or silhouette describe clusters individually but fail to capture the broader relational geometry or multi-hop dependencies between regions.

We address this limitation by introducing **Graph Memory (GM)**, a structured, reusable memory that encodes both representative prototypes and their relations within the embedding space. Each node corresponds to a cluster-level prototype annotated with reliability indicators (e.g., purity, variance, silhouette), while edges capture meaningful interactions between regions, such as geometric proximity or overlap under perturbations. By organizing the embedding space into a connected network of reliable prototypes, GM provides a compact representation that supports both context-aware inference and interpretable reasoning.

GM operates in two stages. In the **memory construction stage**, embeddings are clustered into prototypes, each storing its centroid and reliability metadata. Edges are then established between prototypes according to geometric similarity and structural overlap, forming a concise graph that summarizes the manifold structure of the embedding space. In the **inference stage**, a new query is attached to its most relevant prototypes, and predictions are obtained through

[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

reliability-weighted label propagation over the prototype graph. The resulting ego-graph of the query serves as a localized explanation, highlighting which prototypes supported the decision and how reliable their contributions were.

Conceptually, GM unifies three classical ideas, non-parametric retrieval, prototype-based interpretability, and graph-based semi-supervised reasoning, within a single inductive framework. The resulting method yields calibrated, context-aware, and geometrically smooth predictions, while remaining fully compatible with modern embedding models and scalable non-parametric inference. By transforming the embedding space into a structured relational memory, GM enables efficient decision-making and faithful, region-level explanations of model behavior.

Contributions. This work makes the following contributions:

- We introduce **Graph Memory (GM)**, a structured non-parametric framework that organizes embedding spaces into prototype nodes connected by reliability-aware relations, enabling both inductive inference and faithful explanation.
- We formulate a unified view that generalizes classical methods (kNN, nearest-centroid, and label propagation) under a single prototype-level graph formulation with explicit diffusion and reliability modeling.
- Through experiments on synthetic and real datasets, including breast histopathology (IDC), we demonstrate that GM achieves competitive or superior accuracy and calibration using an order of magnitude fewer samples, confirming its efficiency, robustness, and interpretability.

2 RELATED WORK

This section surveys prior art relevant to embedding-space inference, organized from general to specific. We first contrast parametric and non-parametric classifiers, then review prototype and cluster-based methods, instance retrieval and semi-parametric adapters, and graph-based semi-supervised learning. We next discuss structured relations beyond pure distance (hierarchies and temporal dynamics), summarize efficiency results for serving large non-parametric memories, and close with interpretability, calibration, and cluster-quality assessment. Throughout, we emphasize where existing lines fall short, most notably in lacking a persistent, region-level representation that encodes reliability, ambiguity, and relations among regions in a reusable form.

Parametric *vs.* non-parametric classifiers. Classification based on embeddings is commonly realized with *parametric* heads (e.g., linear-softmax layers) that learn global decision boundaries, or with *non-parametric* schemes (e.g., prototypes, kNN) that rely on neighborhood evidence. While effective, both families rarely maintain an explicit, reusable representation of how *regions* of the space relate, where ambiguity persists, which areas are reliably separated, and how evidence should flow across regions.

Prototype and cluster-based classifiers. Nearest-class-mean and related prototype methods summarize classes via one or a few representatives, while few-shot variants such as Prototypical Networks learn embeddings where class means become discriminative [1]. Prototype-part models like ProtoPNet provide "this-looks-like-that" interpretability within parametric networks [2]. However, prototypes are typically treated in isolation: relations among regions (e.g., persistent cross-class ambiguity, reliability differences between prototypes) are not explicitly modeled, limiting context-aware inference.

Instance retrieval and semi-parametric adapters. Instance-based inference spans classical kNN to Deep kNN (DkNN) [3], which retrieves neighbors across multiple network layers to improve robustness and interpretability. Subsequent semi-parametric adapters integrate retrieval with pretrained encoders at scale, e.g., kNN with frozen vision backbones for privacy-aware continual learning and deletion [4], class-conditional memory that augments parametric logits [5], and kNN-Adapter for black-box LM domain adaptation [6]. While these approaches underscore the value of non-parametric memory, they operate at the *instance* or *layer-activation* level and are tightly coupled to specific deep architectures. In contrast, embedding-level or prototype-level representations can serve as a universal non-parametric head (independent of the encoder or even of neural architectures) allowing consistent reasoning and calibration across arbitrary embedding sources.

Graph-based semi-supervised learning. Graph SSL constructs instance graphs and propagates labels harmonically [7, 8]. Graph Convolutional Networks (GCNs) [9] learn parametric message passing on such graphs. This literature focuses on *propagation rules* given a graph, often in transductive settings. Open questions remain around *graph construction*: which nodes to represent (instances vs. prototypes), which edge semantics to encode (beyond raw similarity), and how to attach new queries in a way that is both scalable and interpretable. Zhou *et al.* [10] further showed that the steady-state solution of a linear propagation process on a similarity graph yields a closed-form classifier $F^* = (I - \alpha S)^{-1} Y$, unifying earlier iterative diffusion and harmonic approaches. This formulation established the link between manifold flatness and global label consistency.

Structured relations beyond distances. Beyond isotropic similarity, hierarchical relations have been modeled with partial-order and box embeddings. Careful objective design and negative sampling are pivotal for high-fidelity taxonomy reconstruction [11]. These results suggest richer edge semantics (e.g., containment or overlap priors) and informed negatives, yet they do not by themselves provide a reusable, region-level memory tailored for classification.

Temporal relations in graphs. Time-decayed line graphs formalize dynamic adjacency with explicit decay kernels over event time, yielding simple and effective temporal edge weighting [12]. Such formulations offer a principled template for incorporating recency into relational signals (e.g., historical ambiguity), though they do not address how to structure prototype-level regions for inference and explanation.

Efficiency and scalability of non-parametric memories. Dhulipala *et al.* propose MUVERA [13] a retrieval mechanism for compressing multi-vector late-interaction retrieval to single-vector MIPS via fixed-dimensional encodings, achieving large latency gains with strong recall. This indicates that large non-parametric memories can be served efficiently with ANN/MIPS backends and product quantization. However, efficiency advances are largely orthogonal to the question of *what* relational structure the memory should encode.

Interpretability, calibration, and cluster quality. Prototype-based interpretability [2] complements post-hoc calibration methods such as temperature scaling [14], Dirichlet calibration [15], and energy-based OOD scoring [16]. Orthogonally, the clustering literature offers a spectrum of *internal* validation measures (e.g., Silhouette [17], Davies-Bouldin [18], Dunn [19], Calinski-Harabasz [20]). A comparative analysis by Liu *et al.* [21] shows these indices capture complementary aspects (compactness, separation, noise sensitivity), with no single metric dominating across conditions. This motivates aggregating multiple, heterogeneous indicators when summarizing region quality and reliability.

Summary of gaps. Prior work contributes essential ingredients: prototypes for interpretability, retrieval for adaptability, semi-parametric adapters for scalability, graph SSL for relational reasoning, hierarchical and temporal modeling for richer relations, and efficient ANN backends for serving. Taken in isolation, however, these lines often (i) treat prototypes or instances without encoding inter-region relations, (ii) lack a persistent, reusable memory over regions, (iii) omit reliability and ambiguity as first-class signals, or (iv) leave graph construction under-specified for scalable, query-time attachment. This leaves open the need for a structured, prototype-level graph memory that jointly encodes reliability, ambiguity, and relations among regions, while remaining compatible with modern retrieval and inference mechanisms, and applicable beyond deep architectures.

3 METHOD

We next introduce **Graph Memory (GM)**, a reusable non-parametric structure that organizes the embedding space into a graph of prototype regions with explicit relational semantics. Each prototype node summarizes a coherent region of the embedding manifold, while edges model contextual similarity and ambiguity between regions. In contrast to standard prototype-based classification (e.g. kNN [22]) that rely solely on distance to isolated representatives, our approach treats inference as *evidence diffusion* over a graph whose nodes carry region-level reliability attributes. The framework is *embedding-agnostic*: it can attach to any fixed feature extractor (neural networks or alternative embedding methods)

without retraining or access to intermediate network layers.

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

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

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

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 [23], 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. This construction captures the local geometric structure of the prototype manifold and provides the topology for evidence diffusion. In practice, these edges are primarily used to quantify contextual coherence, reliability propagation, and regional ambiguity, rather than to modify class boundaries directly.

3.5 Inference by Graph Diffusion

Given a query embedding x_q, its initial affinities to prototypes are computed as

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

where r_c is the reliability weight of prototype c. Rather than classifying by the nearest prototype, inference proceeds by diffusing this activation across the prototype graph. The equilibrium activation vector z is obtained by solving

$$z = (I - \alpha S)^{-1} z_0 = (1 - \alpha) \sum_{t=0}^{\infty} (\alpha S)^t z_0, \quad (4)$$

where $\alpha \in [0, 1)$ controls the diffusion strength and S is the normalized adjacency defined above. This closed form corresponds to the steady-state of the standard label-propagation update ($z^{(t+1)} = (1-\alpha) z_0 + \alpha S z^{(t)}$) (see Zhou *et al.* [10]). Class probabilities are then aggregated over prototypes sharing the same dominant label:

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

The diffusion process smooths activations across connected prototypes, reinforcing consistent regions while revealing areas of cross-class ambiguity. When diffusion is disabled (α=0), inference reduces to an independent voting scheme (e.g. weighted kNN) based solely on prototype reliabilities and distances. As α increases, activations propagate proportionally to prototype reliability and connectivity, producing flatter and more calibrated predictions that reflect the local structure of the embedding manifold.

3.6 Interpretation and Comparison to Baselines

Graph Memory (GM) unifies several classical non-parametric inference schemes within a single framework. Depending on the configuration of prototypes, diffusion strength α, and reliability terms r_c, GM recovers well-known methods as limiting cases:

- **kNN:** when each prototype represents a single instance and diffusion is disabled (α=0), GM reduces to standard k-nearest-neighbor classification.
- **Nearest-centroid:** when all samples of each class collapse to a single prototype and $r_c \equiv 1$, inference corresponds to a nearest-class-mean classifier.
- **Label propagation:** when diffusion is active (α>0) and reliabilities are uniform, the update rule becomes the steady-state solution of a label-propagation system, yielding a semi-parametric graph-based variant.

Unlike these baselines, GM operates on a compact set of region-level prototypes that encode both local geometry and reliability, combining the interpretability of prototype models with the contextual smoothing of graph diffusion. This enables **calibrated, context-aware, and geometrically regular** predictions while maintaining the simplicity and scalability of non-parametric inference. Because it is embedding-agnostic, GM can be applied as a non-parametric

head on top of any pre-trained encoder, from deep neural networks to handcrafted or multimodal feature spaces. By construction, GM provides a principled bridge between instance-level, class-level, and graph-based reasoning, generalizing kNN, nearest-centroid, and label-propagation methods under a single relational formulation. Furthermore, GM offers localized, prototype-level explanations that reveal which regions support a given decision and how reliable their contributions are, extending prior work on dimensionality reduction [24], supervised decision-boundary visualization [25], and stability analysis of decision maps [26].

Role of edges. While prototype activations alone often suffice for accurate predictions, the edge structure provides a complementary view of the embedding manifold. Connectivity patterns among prototypes reveal where regions are coherent, ambiguous, or sparsely supported, which enables outlier detection and visual explanations of class overlap. Thus, GM distinguishes between local evidence (nodes) and contextual reliability (edges), yielding an interpretable non-parametric memory that can explain and calibrate predictions.

4 EXPERIMENTAL EVALUATION

We evaluate Graph Memory (GM) on both synthetic and real binary classification datasets to analyze its behavior, calibration, and interpretability. Experiments progress from controlled two-dimensional *synthetic data* (Sec. 4.2) (which allow direct visualization of prototype structure and decision boundaries) to high-dimensional real data from breast histopathology (IDC; Sec. 4.3). Unless otherwise stated, all methods operate on identical frozen embeddings, sharing the same train/test splits and preprocessing. Hyperparameters are selected on validation data via grid search for discrete parameters and short sweeps for continuous ones. Our experimental code is available at https://github.com/arturandre/memory-graph-prototypes

Metrics. We report Top-1 accuracy and negative log-likelihood (NLL) as quantitative performance and calibration measures. To assess the geometric regularity of decision functions, we additionally compute the *Dirichlet energy*, which quantifies how rapidly class probabilities vary across the embedding space.

For two-dimensional datasets, the energy is estimated as the mean squared gradient magnitude of the class-1 probability field over a dense grid:

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

where lower values indicate smoother and more stable decision transitions. Gradient-magnitude maps are shown side-by-side for kNN, Label Spreading, and GM to visualize boundary regularity.

For higher-dimensional embeddings, the same principle is measured via the *graph Dirichlet energy*, a standard notion of smoothness in signal processing on graphs [27]:

$$E(f) \;=\; \tfrac{1}{2|\mathcal{E}|}\sum_{(i,j)\in\mathcal{E}} w_{ij}\left(f_i - f_j\right)^2 \qquad (5)$$

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

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 against representative parametric and non-parametric models:

- **Linear probe:** multinomial logistic regression.
- **kNN:** instance-level k-nearest neighbors.
- **Budget-kNN:** kNN trained on a stratified subset whose size matches GM's prototype budget ($|\text{train}|{=}K$).
- **Label Spreading (LS):** scikit-learn transductive label spreading on an instance graph with n-neighbors.

4.2 Synthetic datasets

We first evaluate our approach on the *moons* and *circles* datasets from scikit-learn [28] (v1.7.2), each with $n{=}4000$ samples and a 50/50 stratified train/test split. For long-tail stress tests, we downsample the training data to yield a class imbalance of 8:1, with 1,000 samples from the majority class and 125 from the minority.

Graph Memory (GM) builds a joint prototype set across classes via K-means clustering, using $K = \max(32, \min(120, n/10)) = 120$ for the balanced case and K=112 for the imbalanced setting (n=1125). Each prototype connects to its k_{graph}=10 nearest prototype neighbors, with diffusion strength α=0.5, inverse-radius β=0.1, and attachment degree attach-k=8. Fig. 1 shows representative decision regions, and Figure 2 illustrates gradient magnitude maps indicating the flatness of the learned decision boundaries.

Table 1 summarizes the quantitative comparison across methods. All models are evaluated over five runs, reporting mean (std) for accuracy (Acc ↑) and negative log-likelihood (NLL ↓). **Graph Memory (GM)** attains accuracy comparable to kNN and Label Spreading while achieving markedly better calibration (lower NLL). Unlike kNN, GM relies on a compact prototype memory (only about 10% of the training samples) and unlike Label Spreading, it remains fully **inductive**, requiring no access to test data at inference. GM maintains its calibration and robustness even under strong class imbalance. For reference, kNN-budget(P=112) denotes kNN restricted to the same number of points as GM, illustrating how naive subsampling severely degrades performance in the long-tail setting, whereas GM remains competitive. The Linear model is included only as a reference, since these datasets are non-linearly separable.

Table 2 reports the *graph Dirichlet energy* (as defined in Eq. 5), which measures the average squared gradient magnitude of the class-probability field. Lower values indicate smoother and more stable decision regions. GM consistently achieves lower energy, reducing the graph Dirichlet energy by **3040%** relative to kNN and by over **94%** compared to Label Spreading. These results confirm that GM produces more regular and robust decision surfaces while maintaining high accuracy and calibration.

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)

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

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 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 kNN and over **94%** below Label Spreading) indicating smoother, more reliable boundaries while preserving accuracy and calibration.

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] [29, 30]. Each sample is a 50×50 histopathology patch labeled as benign or malignant (IDC). Representative examples of benign and malignant tissue patches are shown in Figures 3 and 4, respectively.

Frozen 512-D embeddings are extracted using a ResNet18 encoder trained for five epochs with standard data augmentations, reaching a validation accuracy of 0.885. This ensures that the backbone provides sufficiently discriminative representations, allowing the evaluation to focus on the inference behavior of Graph Memory rather than feature quality. From these embeddings, GM is instantiated with the same hyperparameters used in the synthetic experiments, except for a fixed training and testing split of 2000 samples each and a compact prototype memory of K=32. Each prototype connects to its k_{graph}=10 nearest

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

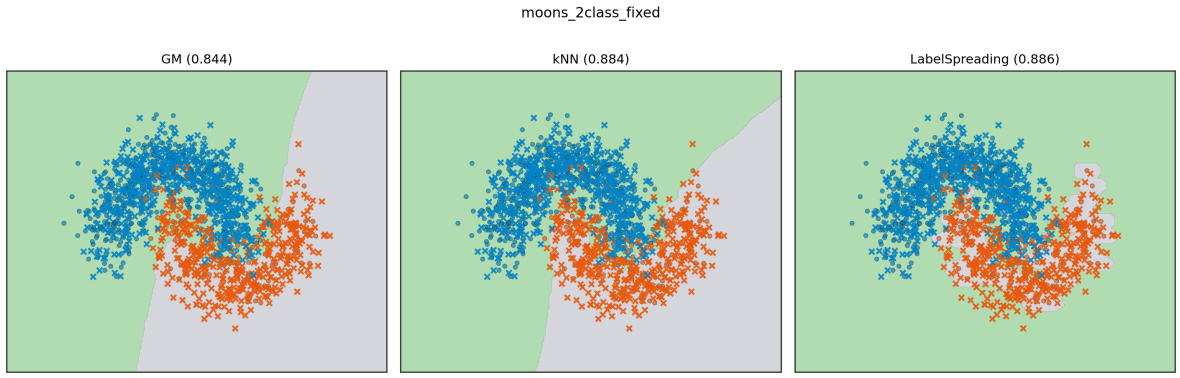

Figure 1: Decision regions on synthetic data. GM yields flatter, reliability-weighted boundaries while retaining prototype-level interpretability.

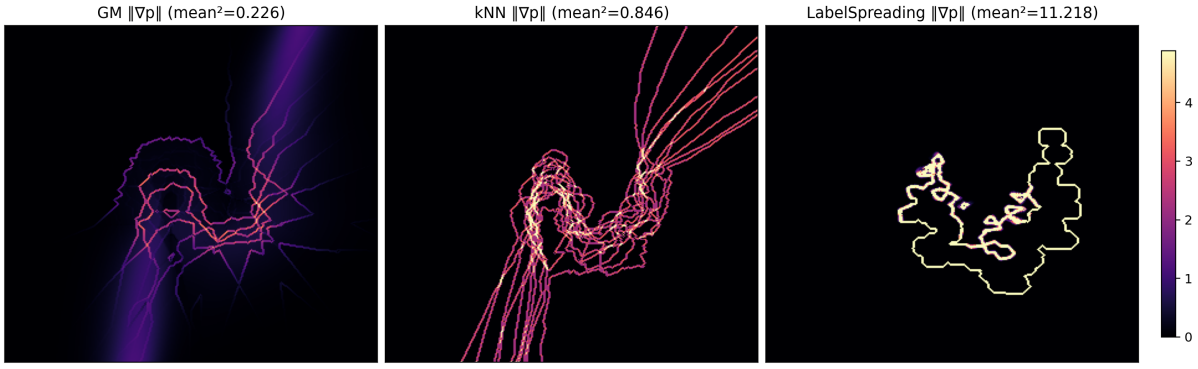

Figure 2: **Dirichlet energy maps** (visualized as gradient-magnitude fields, $\|\nabla p(y{=}1 \mid x)\|$). Lower values indicate smoother and more stable decision transitions. GM produces lower-energy, reliability-weighted boundaries and avoids the sample-hugging artifacts observed in Label Spreading and kNN, particularly near class interfaces.

neighbors, with diffusion strength α=0.5, inverse-radius β=0.1, and attachment degree attach-k=8.

We report accuracy, negative log-likelihood (NLL), and graph Dirichlet energy (Equation 5) as the main evaluation metrics. GM delineates smooth and reliable decision boundaries while retaining prototype-level interpretability.

Table 3 summarizes the quantitative comparison. **Graph Memory (GM)** attains the highest accuracy while maintaining strong calibration, despite using only **32 prototypes**, a compact and efficient representation of the entire dataset. The Linear classifier achieves a slightly lower NLL (0.397 vs. 0.420) due to its parametric optimization, but with comparable accuracy (0.836 vs. 0.845), indicating that GM reaches near-parametric calibration while remaining fully non-parametric and interpretable. All methods except the kNN-budget baseline exhibit negligible variance across five runs (std $\approx$ 0), confirming stable convergence and consistent behavior under this setup.

Table 4 further compares the *graph Dirichlet energy*, which quantifies the smoothness of decision boundaries. GM achieves the lowest energy among all methods (even in this fully inductive setting) indicating smoother and more stable decision functions than both kNN and Label Spreading.

Figure 3: Representative benign tissue patches from the IDC dataset.

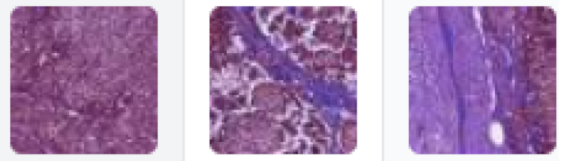

Figure 4: Representative malignant tissue patches from the IDC dataset.

Method	Acc ↑	NLL ↓
GM(P=32)	**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 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	graph Dirichlet energy
GM (P=32)	0.291 (0.000)
kNN	0.321 (0.000)
Label Spreading	0.344 (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 kNN and Label Spreading while maintaining negligible variance across runs.

5 Discussion

Across both synthetic and real-world datasets, Graph Memory (GM) demonstrates that embedding-level reasoning can benefit from structured, prototype-based graph representations. By organizing the embedding space into reliability-weighted regions connected through local relations, GM unifies key aspects of classical non-parametric methods (instance retrieval, prototype abstraction, and label propagation) within a single inductive framework.

Empirically, GM achieves accuracy on par with or above kNN and Label Spreading while delivering markedly better calibration and smoother decision boundaries, as evidenced by lower graph Dirichlet energy across all tested settings. Under matched memory budgets, GM consistently outperforms instance-based baselines (Budget-kNN), highlighting the representational advantage of structured region prototypes over randomly selected instance caches. Moreover, GM's compact memory (often less than 10% of the training set) achieves comparable or superior results, confirming its efficiency and scalability.

Label Spreading remains a strong transductive baseline on clean manifolds, but its sample-hugging boundaries lead to higher energy values and degraded calibration. In contrast, GM generalizes to unseen samples without access to test data, maintaining robust calibration even in imbalanced or high-dimensional scenarios such as breast histopathology (IDC). These results suggest that modeling regional reliability and contextual diffusion provides a principled way to bridge local evidence and global consistency.

Beyond performance, GM provides interpretability by explicitly separating node-level evidence (prototypes) from edge-level context (relations). This separation exposes which regions contribute most to a prediction and how contextual dependencies shape the decision surface, offering a transparent view of the inference process. Such structured interpretability positions GM as a strong inductive counterpart to traditional graph-based semi-supervised learning, while also serving as a foundation for explainable and reliability-aware non-parametric inference.

Graph-structured prototype memories align with a broader trajectory of works emphasizing the integration of geometric, semantic, and statistical information for reliable and interpretable decision-making. Earlier multimodal approaches combined deep learning with metadata and data-integration pipelines to improve calibration and robustness in complex visual domains such as urban imagery and vegetation monitoring [31, 32]. When these graph structures are enriched with additional metadata and statistical descriptors, they not only enhance compression and calibration but also foster interpretability, offering a view of how each prototype aggregates and contextualizes evidence. Recent research on interactive semantic mapping and fused embeddings [33, 34] further demonstrates that coupling embeddings with user-steerable or multi-modal attributes leads to more interpretable and controllable models. In this context, Graph Memory can be seen as a generalization of these ideas: by explicitly encoding reliability and relational metadata within a prototype graph, it provides a compact yet explainable representation that supports both quantitative precision and qualitative insight.

6 Conclusion

We introduced **Graph Memory (GM)**, a structured non-parametric framework that augments embedding-based inference with prototype-level relations and reliability modeling. GM generalizes classical methods such as kNN, nearest-centroid, and label propagation under a unified graph formulation, enabling calibrated, region-aware, and geometrically smooth predictions. Experiments on synthetic and real datasets show

that GM delivers high accuracy, strong calibration, and low graph Dirichlet energy while operating on a compact memory of prototypes.

Beyond classification, the concept of a reusable, reliability-annotated graph memory opens promising directions for explainable AI and continual learning. Future work includes extending GM to multi-modal embeddings, temporal dynamics, and large-scale retrieval systems, where structured regional memory could serve as a bridge between symbolic knowledge graphs and dense representation learning.

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).

REFERENCES

[1] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

[2] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Barnett, Cynthia Rudin, and Jonathan Su. This looks like that: Deep learning for interpretable image recognition. In *NeurIPS*, 2019.

[3] Nicolas Papernot and Patrick McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. In *International Conference on Learning Representations*, 2018.

[4] Sebastian Doerrich, Tobias Archut, Francesco Di Salvo, and Christian Ledig. Integrating knn with foundation models for adaptable and privacy-aware image classification. *arXiv preprint arXiv:2402.12500*, 2024.

[5] Himanshu Bhardwaj, Danish Pruthi, and Graham Neubig. knn-cm: Semi-parametric classifiers via *k*nn-based class-conditional memory. In *Findings of the Association for Computational Linguistics: EMNLP*, 2023.

[6] Yangsibo Huang, Daogao Liu, Zexuan Zhong, Weijia Shi, and Yin Tat Lee. knn-adapter: Efficient domain adaptation for black-box language models. In *arXiv preprint arXiv:2302.10879*, 2023.

[7] Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, 2003.

[8] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

[9] Thomas Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

[10] Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. *Advances in neural information processing systems*, 16, 2003.

[11] Alyssa Lees, Chris Welty, Shubin Zhao, Jacek Korycki, and Sara Mc Carthy. Embedding semantic taxonomies. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1279–1291, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[12] Sudhanshu Chanpuriya, Ryan A. Rossi, Sungchul Kim, Tong Yu, Jane Hoffswell, Nedim Lipka, Shunan Guo, and Cameron Musco. Direct embedding of temporal network edges via time-decayed line graphs, 2022.

[13] Laxman Dhulipala, Majid Hadian, Rajesh Jayaram, Jason Lee, and Vahab Mirrokni. Muvera: Multi-vector retrieval via fixed dimensional encodings, 2024.

[14] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

[15] Meelis Kull, Miquel Perello-Nieto, Miquel Kängsepp, Telmo Silva Filho, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. *NeurIPS*, 2019.

[16] Weitang Liu, Xiaoyun Wang, John D Owens, and Chelsea Finn. Energy-based out-of-distribution detection. In *NeurIPS*, 2020.

[17] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

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

[19] Joseph C. Dunn. Well-separated clusters and optimal fuzzy partitions. 1974.

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

[21] Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu. Understanding of internal clustering validation measures. In *2010 IEEE International Conference on Data Mining*, pages 911–916, 2010.

[22] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[23] Peter Rousseeuw and Christophe Croux. Alternatives to median absolute deviation. *Journal of the American Statistical Association*, 88:1273 – 1283, 12 1993.

[24] Artur André AM Oliveira, Mateus Espadoto, Roberto Hirata Jr, Nina ST Hirata, and Alexandru C Telea. Improving self-supervised dimensionality reduction: Exploring hyperparameters and pseudo-labeling strategies. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics*, pages 135–161. Springer, 2021.

[25] Artur André Almeida de Macedo Oliveira, Mateus Espadoto, Roberto Hirata Júnior, and Alexandru Cristian Telea. Sdbm: supervised decision boundary maps for machine learning classifiers. In *Proceedings*, 2022.

[26] Artur AAM Oliveira, Mateus Espadoto, Roberto Hirata Jr, and Alexandru C Telea. Stability analysis of supervised decision boundary maps. *SN Computer Science*, 4(3):226, 2023.

[27] David I Shuman, Sunil K. Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. 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, 2013.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

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

[30] Angel Cruz-Roa, Ajay Basavanhally, Fabio González, Hannah Gilmore, Michael Feldman, Shridar Ganesan, Natalie Shih, John Tomaszewski, and Anant Madabhushi. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In *Medical imaging 2014: Digital pathology*, volume 9041, page 904103. SPIE, 2014.

[31] Artur Andre Almeida de Macedo Oliveira. Overcoming challenging crban images: deep learning and data integration methods for detecting trees entangled with power lines. 2023.

[32] Artur André Oliveira and Roberto Hirata Jr. Deep learning and data integration for detecting trees entangled with utility lines. *CLEI Electronic Journal*, 28(6):3–1, 2025.

[33] Artur André Oliveira, Mateus Espadoto, Roberto Hirata Jr, Roberto M Cesar Jr, and Alex C Telea. Creating user-steerable projections with interactive semantic mapping. *arXiv preprint arXiv:2506.15479*, 2025.

[34] Artur A. Oliveira, Mateus Espadoto, Roberto Hirata Jr., and Roberto M. Cesar Jr. Improving image classification tasks using fused embeddings and multimodal models. In *Proceedings of VISIGRAPP (2: VISAPP) 2025*, volume 232, pages 232–241. SciTePress, 2025.