TopoAgent: An Agentic Framework for Automated Topology Learning in Medical Imaging

Guangyu Meng[1], Pengfei Gu[2†], Xueyang Li[1], Yiyu Shi[1], Erin Wolf Chambers[1†], and Danny Z. Chen[1†]

[1] Dept. of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, USA
[2] Dept. of Computer Science, The University of Texas Rio Grande Valley, Edinburg, TX, USA

{gmeng, xli34, Yiyu.Shi.31, echambe2, dchen}@nd.edu, pengfei.gu01@utrgv.edu

Abstract. Topological data analysis (TDA), particularly persistent homology (PH), captures geometric structural properties in medical images (e.g., connected components, loops, shape characteristics), which conventional pixel-level deep learning approaches often neglect. While many topological descriptors are known for converting persistence diagrams (PDs) or raw images into topological feature vectors, existing methods mostly default to a single fixed descriptor (e.g., persistence images), leaving the diversity of topological representations largely unexplored. To the best of our knowledge, there is no known large language model (LLM)-based agentic framework that can automatically determine the most suitable topological descriptors for a given image dataset and produce the corresponding topological feature vectors for downstream tasks. To fill this gap, we propose **TopoAgent**, an LLM-based agentic framework that automates topology learning for medical image analysis. TopoAgent operates through a Perception–Reasoning–Action–Reflection loop supported by 21 domain-specific tools and dual memory that accumulates experience across runs. Its skill set is distilled from systematic evaluation of 15 topological descriptors across 26 datasets with six classifiers. TopoAgent analyzes input images and their topological characteristics, reasons about which topological descriptors best suit the input, and determines the optimal descriptor and its configuration, all without task-specific training. To evaluate TopoAgent, we introduce **TopoBenchmark**, a frozen benchmark of 113,182 samples from 26 medical image datasets spanning five object types: cells, glands and lumens, organ shapes, vessel trees, and surface lesions. Experiments show that TopoAgent obtains 68.21% average balanced accuracy, outperforming the strongest baseline by 9.32% and general-purpose LLMs equipped with the same tools by over 21%.

Keywords: Topological Data Analysis · Persistent Homology · Topology Learning · Medical Image Analysis · LLM-based Agent · Benchmark

† Corresponding authors.

1 Introduction

Persistent homology (PH), a central tool in topological data analysis (TDA), captures geometric structural properties in images that conventional pixel-level deep learning (DL) approaches often neglect [1,10,16,36,41]. In medical imaging, this topological perspective has proven valuable for grading cancer [33,57], predicting breast cancer survival [50], analyzing retinal vasculature [4], and classifying tissue morphology [23,24], with growing adoption in other visual domains such as shape analysis and materials science [52]. To make these topological insights actionable for downstream tasks, researchers rely on *topological descriptors*: methods that convert persistence diagrams (PDs) or raw images into fixed-length topological feature vectors [5, 52]. A rich family of such descriptors is known, including persistence images [2], persistence landscape [9], Euler characteristic transform [54], and Minkowski functional [40], among others (see Supplementary for the complete descriptor taxonomy). However, each descriptor encodes distinct geometric characteristics and entails specific mathematical properties, hyperparameters, and failure modes. No single descriptor is universally effective across all datasets [5] (e.g., see Fig. 1(a)). Thus, determining effective descriptors for a given image dataset is an overwhelming challenge that demands a great deal of expertise and effort.

This determination is challenging because it depends on the interaction between dataset-specific morphological characteristics (e.g., connected components, loops, cavities) and each descriptor's mathematical properties. In practice, users must manually analyze dataset properties, choose among candidate descriptors, tune hyperparameters, and produce topological feature vectors for downstream tasks. With a large descriptor library and continuous parameter spaces, exhaustive evaluation is prohibitive at the per-image level, making this trial-and-error process time-consuming, expertise-dependent, and difficult to standardize.

Large language model (LLM)-based agentic frameworks offer a "natural" solution. Modern agents integrate reasoning, tool use, memory, and iterative refinement [49, 63], and medical AI agents have begun orchestrating domain-specific tools for clinical analysis [18, 35, 59, 61]. Since descriptor determination requires both high-level reasoning about data morphology and low-level computation (e.g., PH and descriptor extraction), it aligns naturally with this paradigm. However, no existing agentic framework automatically computes PH, reasons about topological structures, and produces topological feature vectors for downstream tasks. General-purpose LLMs can describe topology in natural language but cannot compute it without external tool integration [37], and even with tools, they lack the empirical grounding to verify their determinations (e.g., see Fig. 1(b)).

Automating descriptor determination also requires systematic evaluation, but prior studies typically assess a small number of descriptors on limited datasets with varying preprocessing, classifiers, and evaluation metrics [5], making cross-study comparisons unreliable. No standardized benchmark spans diverse medical image morphologies with consistent criteria, leaving the community without clear guidance on when specific topological descriptors excel or fail.

(a) Accuracy of different topological descriptors on five datasets

Dataset	PS [5]	PI [2]	TF [46]	ATOL [48]	MK [40]	LBP [44]	Best
BloodMNIST [62]	96.75	90.63	**97.92**	97.64	97.75	90.03	TF
BreakHis [51]	**88.76**	67.92	79.64	87.03	83.91	77.02	PS
BrainTumor [14]	95.10	83.49	95.12	**96.16**	94.71	94.98	ATOL
ISIC2019 [15]	36.14	28.17	33.31	34.05	**36.94**	29.66	MK
APTOS2019 [7]	55.03	39.65	53.84	51.83	52.06	**57.74**	LBP

(b) Reasoning comparison on a BreakHis image

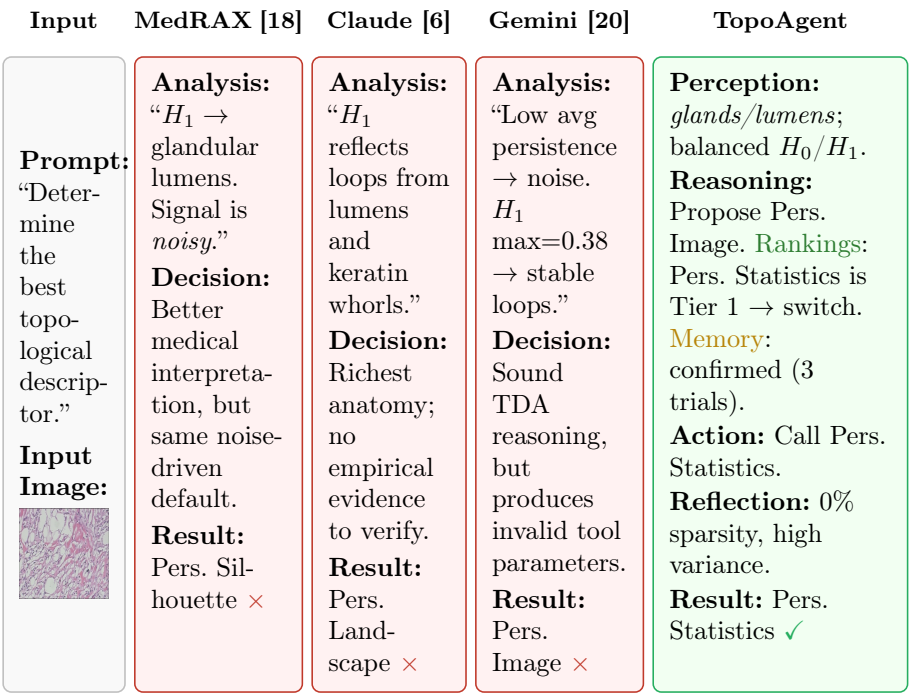

Fig. 1: (a) Balanced accuracy (%) of 6 descriptors (covering the top-3 per object type) on 5 datasets. PS = persistence statistics, PI = persistence image, TF = template function, ATOL = automatic topologically-oriented learning, MK = Minkowski functional, LBP = local binary pattern. The best descriptor (bold) differs across 5 datasets, confirming no single descriptor is best for all. **(b)** Reasoning comparison on a BreakHis image (Pers. = persistence). General-purpose LLMs with the same tools all select suboptimal descriptors; GPT-4o [28] (MedRAX's backbone) reaches similar result without medical insight. TopoAgent determines persistence statistics for glands/lumens.

To address these gaps, we propose **TopoAgent** and **TopoBenchmark**. TopoAgent is an LLM-based agentic framework that automates topology learning for medical images by determining the most suitable topological descriptors for an input dataset. Given a raw image and a task prompt, TopoAgent oper-

ates through a Perception–Reasoning–Action–Reflection (PRAR) loop without task-specific training. Perception characterizes the image's topological and visual properties using perception tools. Reasoning first formulates a descriptor proposal by matching the observed characteristics based on descriptor properties, and then integrates empirical evidence from a distilled skill set and dual memory to determine the final descriptor and parameters. Action executes the determined descriptor using descriptor tools to produce a topological feature vector. Reflection validates the output quality and, on failure, records a diagnosis in long-term memory and triggers retry, enabling the agent to self-correct rather than committing to a single-shot decision. TopoBenchmark is a standardized evaluation benchmark comprising 26 datasets that broadly cover publicly available 2D medical imaging scenes, organized into five object types that span the major morphological categories in medical images: *cells*, *glands and lumens*, *organ shapes*, *vessel trees*, and *surface lesions*. It provides an empirical basis from which the skill set is distilled and a frozen testbed of 113,182 samples with convergence-based per-dataset sample sizing for consistent evaluation.

Our contributions are threefold. (1) We introduce TopoAgent, the first agentic framework that bridges LLM-based reasoning and topological data analysis, enabling automated topology learning from raw medical images with full reasoning traces. (2) We construct TopoBenchmark, the first standardized benchmark for evaluating topological descriptors across diverse medical image morphologies with consistent criteria. (3) We demonstrate that adaptive, per-image descriptor determination outperforms the strongest baseline by 9.32% and general-purpose LLMs equipped with the same tools by over 21% in average balanced accuracy. Our code is publicly available at `https://github.com/gm3g11/TopoAgent`.

2 Background and Related Work

2.1 Topological Descriptors

PH tracks the appearance and disappearance of topological features across a filtration of the input image I. For 2D medical images, we apply a sublevel-set filtration on pixel intensities via cubical complexes, yielding a PD $\mathrm{Dgm}_k(I)$: a multiset of birth-death pairs $\{(b_i, d_i)\}$ recording when each k-dimensional feature appears (b_i) and vanishes (d_i). Here, $k = 0$ captures connected components and $k = 1$ captures loops. The persistence $d_i - b_i$ measures each feature's significance across scales [10,16]. To make PDs compatible with machine learning, topological descriptors convert them into fixed-length vectors. These fall into two families: PH-based descriptors $\varphi_d(\mathrm{Dgm}(I); \theta) \to \mathbb{R}^{n_d}$ that operate on PDs (e.g., persistence images [2], persistence landscape [9], Betti curve [55]), and image-based descriptors $\varphi_d(I; \theta) \to \mathbb{R}^{n_d}$ that extract topological or geometric features directly from the image, bypassing PH entirely (e.g., Euler characteristic transform [54], Minkowski functional [40]). Each descriptor encodes different structural properties with distinct mathematical properties, and no single descriptor is universally effective across all datasets [5]. We refer readers to [5, 52] for comprehensive surveys of topological descriptors. From these surveys and literature [3, 12, 43, 65],

we identify 15 descriptors that are widely adopted, mathematically well-founded, and applicable to 2D medical images, forming the descriptor library $\mathcal{D}$ used throughout this work. Recent work has integrated topological features into DL pipelines for image classification through differentiable topological layers [11,31], notably PHG-Net [45], which fuses learned persistence features with CNN and Transformer backbones. These approaches learn a fixed vectorization end-to-end rather than determining which descriptor suits a given input.

2.2 LLM-based Agentic Frameworks

LLM-based agents extend LLMs beyond text generation by equipping them with tools, memory, and structured reasoning loops. ReAct [63] introduced the paradigm of interleaving chain-of-thought reasoning with tool calls, enabling models to ground their reasoning in external observations. Reflexion [49] added self-reflection, where agents learn from past failures stored in verbal memory to improve subsequent attempts. Voyager [56] demonstrated open-ended skill acquisition through a growing library of reusable behaviors. These ideas have been synthesized into the PRAR architecture pattern [30, 42, 60], which our TopoAgent adopts as the backbone. More broadly, skills in agentic AI are reusable knowledge modules that encode domain expertise to augment an agent's reasoning [58, 60]; they range from executable code snippets to structured decision rules. In medical AI, agentic systems have begun orchestrating domain-specific tools: MedRAX [18] coordinates radiological analysis tools for chest X-ray reasoning, MMedAgent [35] routes queries across multi-modal medical tools, and MedAgent-Pro [59] chains diagnostic workflows with tool-augmented reasoning. However, no such frameworks compute PH, reason about topological structures, and produce topological feature vectors. An alternative to agentic determination is automated machine learning (AutoML) [19,29], which searches parameter spaces via black-box optimization. However, AutoML treats descriptor choice as a hyperparameter to tune rather than a reasoning problem that requires understanding of the interaction between data topology and descriptor properties, and thus cannot generalize from prior experience to unseen inputs without rerunning the search.

3 Method

3.1 Framework Overview

Given a medical image I and a task prompt, TopoAgent determines the most suitable topological descriptor $d^* \in \mathcal{D}$ from a library of $|\mathcal{D}| = 15$ descriptors, configures its parameters θ^*, and produces a topological feature vector $\mathbf{f} \in \mathbb{R}^{n_d}$ along with a reasoning trace $\mathcal{R}$ documenting all decisions. It uses the PRAR architecture [30, 58, 60], extended with tools, a distilled skill set $\mathcal{S}$, and dual memory (M_s, M_l). The skill set $\mathcal{S}$ comprises three components: descriptor properties $\mathcal{S}_{\text{prop}}$ encoding qualitative knowledge, empirical evidence $\mathcal{S}_{\text{rank}}$ providing

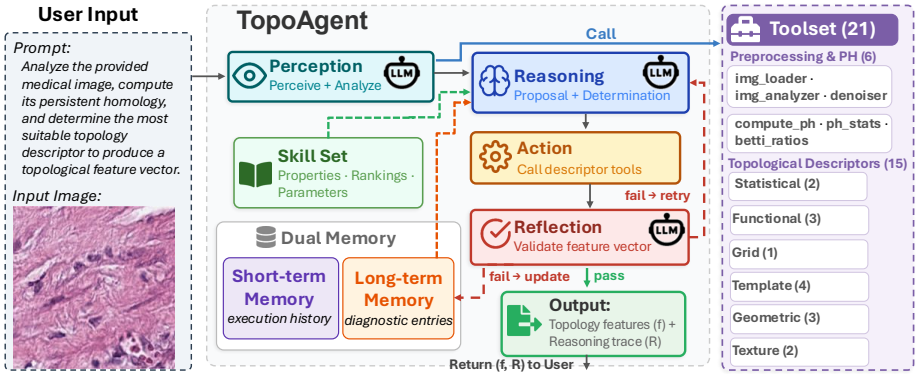

Fig. 2: An overview of the TopoAgent framework with four phases: **Perception** identifies the object type and analyzes persistent homology statistics; **Reasoning** proposes and determines the descriptor and parameters; **Action** calls the descriptor tools; **Reflection** validates the feature vector. The skill set $\mathcal{S}$ feeds descriptor properties and tiered rankings with parameters to Reasoning (dashed green). Short-term memory M_s tracks execution history within a run; long-term memory M_l supplies diagnostic entries to Reasoning (dashed orange) and is updated on failure.

tiered rankings and reasoning chains, and validated parameters $\mathcal{S}_{\text{param}}$ specifying optimal configurations per object type. Short-term memory M_s records tool invocations within each run; long-term memory M_l keeps diagnostic entries across runs. Fig. 2 shows the workflow. We present the four-phase pipeline in §3.2 and the supporting components in §3.3.

3.2 TopoAgent Pipeline

Perception. Before reasoning about descriptor suitability, the agent must first understand the image: its imaging modality, tissue or anatomical structures, and topological characteristics. Perception invokes six perception tools deterministically, producing three outputs. (a) The PH profile $h(I)$ summarizes the topological characteristics: birth-death pair counts per homology dimension, average persistence, Betti ratio β_1/β_0 (the ratio of loop count to component count), and per-dimension persistence distributions. (b) Visual statistics $v(I)$ capture image-level properties such as signal-to-noise ratio, contrast, and edge density. (c) The object type $o \in \mathcal{O} = \{\textit{cells}, \textit{glands/lumens}, \textit{organ shapes}, \textit{vessel trees}, \textit{surface lesions}\}$ is identified from the image content. Deterministic tool execution ensures reproducible PH computation. The LLM's role in Perception is interpretive: it jointly reasons on the raw image and the computed PH to identify o. The object type o is critical because it indexes the skill set: empirical evidence $\mathcal{S}_{\text{rank}}$ and validated parameters $\mathcal{S}_{\text{param}}$ are all organized per object type.

Both topological and visual information are necessary for reliable object-type identification. The PH profile quantifies topological characteristics that directly inform descriptor choice, such as the relative prevalence of connected components versus loops and their persistence distributions. However, the PH profile

alone cannot reliably distinguish all object types. For example, both dense cell populations and glandular tissues with fragmented lumen boundaries can give somewhat similar H_0-dominated PH profiles, but they require fundamentally different descriptors due to distinct underlying morphology. Visual grounding resolves such ambiguities: jointly observing the image and its PH substantially outperforms both PH-profile-only and vision-only identification (Table 2), confirming that the two information sources provide complementary evidence.

Reasoning. Reasoning is the central phase of the pipeline, responsible for both proposing and determining the descriptor configuration. It addresses the core challenge of the framework: determining an effective topological descriptor requires understanding of the interaction between the image's topological characteristics and each descriptor's mathematical properties. To handle this, Reasoning takes two steps with asymmetric information access.

In the *proposal step*, the agent receives the Perception outputs $(h(I), v(I), o)$ together with $\mathcal{S}_{\text{prop}}$, which encodes the mathematical definition, strengths, weaknesses, and intended use of each descriptor. Crucially, $\mathcal{S}_{\text{rank}}$ is provided in stripped form: the agent sees reasoning patterns that describe how PH characteristics relate to descriptor suitability, but not the descriptor recommendations or tiered rankings themselves. The agent reasons about which descriptor's mathematical properties best match the observed topological characteristics and proposes a candidate (d_p, θ_p) with preliminary parameters. In the *determination step*, the agent integrates the full $\mathcal{S}_{\text{rank}}$ (including tiered rankings, descriptor recommendations, and threshold-based PH signal definitions) and long-term memory M_l (outcomes from prior runs on the current dataset). Weighing the proposal against this empirical evidence, the agent arrives at one of three decisions: it retains (d_p, θ_p) when the PH profile provides strong evidence despite a low ranking, defers to $\mathcal{S}_{\text{rank}}$ when the proposal is weakly supported, or adopts a correction from M_l that overrides both.

This two-step design addresses anchoring bias, the tendency of LLMs to fixate on the first salient information that they receive rather than reasoning independently. When the full $\mathcal{S}_{\text{rank}}$ is visible during proposal formation, the LLM determines the top-ranked descriptor in 92.1% of 520 held-out test images (20 per dataset in 26 datasets), even when explicitly prompted to "refer to but not rely on" the rankings. This reduces the agent into a lookup table. Providing only stripped reasoning patterns during the proposal step reduces the top-ranked agreement rate to 61.3%: of these, 38.8% are cases where the agent independently arrived at the same determination, while 22.5% are cases where $\mathcal{S}_{\text{rank}}$ corrected the proposal during determination (more details in the Supplementary).

Action. Action executes the determined descriptor d^* with parameters θ^* by invoking the corresponding descriptor tools, producing a topological feature vector $\mathbf{f} \in \mathbb{R}^{n_d}$. The parameters are drawn from the ranges validated during skill set construction for the identified object type o rather than generated by the LLM, ensuring reproducible and correctly configured tool calls.

Reflection. A single pass through Reasoning and Action does not guarantee a suitable output: the determined descriptor may produce an unsuitable feature

vector due to parameter misconfiguration (e.g., overly fine resolution causing excessive sparsity), a mismatch with the image's topological characteristics, or edge-case PH characteristics not anticipated by $\mathcal{S}$. Reflection addresses this by having the LLM assess the quality of $\mathbf{f}$ based on summary statistics (sparsity, variance, kurtosis, skewness, dynamic range, and informative-feature ratio) together with descriptor-specific reference ranges from $\mathcal{S}$ and diagnostic entries from M_l. Crucially, reference ranges are provided as context, not as hard thresholds: what constitutes acceptable quality varies across descriptors and tissue types. For instance, high sparsity in a persistence image may be appropriate for sparse PH but indicates failure on dense glandular tissue; LBP histograms naturally exhibit high kurtosis that fixed thresholds would flag as anomalous. If the LLM diagnoses a quality failure, it returns a structured correction (e.g., "switch to a lower-dimensional descriptor"), recorded in M_l along with the failed descriptor and diagnosed cause, then retries from the determination step of Reasoning. A maximum of two retries is permitted within $t_{\max}$; if no passing result is obtained, the agent falls back to the top-ranked descriptor for o from $\mathcal{S}_{\text{rank}}$. Reflection validates feature vector quality, not downstream classification accuracy, since class labels are unavailable at inference time. On success, TopoAgent returns $\mathbf{f}$ together with a reasoning trace $\mathcal{R} = \{h(I), v(I), o, d_p, d^*, \theta^*, q\}$, where q denotes the quality diagnosis.

3.3 TopoAgent Components

Toolset. The agent accesses 21 domain-specific tools organized into two groups. 6 *perception tools* support the Perception phase: image loading, quality analysis, denoising, PH computation via GUDHI [39] with cubical complex filtration, PH profile extraction, and Betti ratio computation. 15 *descriptor tools* support the Action phase: 10 PH-derived descriptors that vectorize persistence diagrams (e.g., persistence image [2], persistence landscape [9], ATOL [48]) and 5 image-based descriptors that capture geometric properties without explicit PH computation (e.g., Minkowski functional [40], Euler characteristic transform [54], LBP texture [44]); see Supplementary for the complete taxonomy. Each tool exposes configurable parameters whose optimal values per object type are determined during skill set construction. Implementing descriptors as pre-built, parameterized tools ensures that the LLM reasons about descriptor suitability rather than generating computation code.

Skill Set. To determine suitable descriptors, TopoAgent needs domain expertise that LLMs lack: understanding of each descriptor's mathematical properties, empirical evidence of descriptor effectiveness per object type, and appropriate parameter configurations. We encode this expertise in a static skill set $\mathcal{S}$, distilled offline from systematic evaluation on TopoBenchmark (§3.4): descriptor properties $\mathcal{S}_{\text{prop}}$, empirical evidence $\mathcal{S}_{\text{rank}}$, and validated parameters $\mathcal{S}_{\text{param}}$.

As shown in Fig. 3, $\mathcal{S}_{\text{prop}}$ encodes qualitative knowledge for the proposal step: the mathematical definition, strengths, weaknesses, and intended use of each descriptor, along with parameter reasoning heuristics that guide the LLM toward appropriate configurations given the observed PH profile. $\mathcal{S}_{\text{rank}}$ provides

empirical evidence for the determination step: per-object-type descriptor rankings presented as tiers rather than exact accuracy values, reasoning chains that map PH profile patterns to descriptor recommendations, and threshold-based PH signal definitions. As detailed in §3.2, the proposal step receives reasoning patterns without descriptor recommendations, while the determination step receives the full evidence including tiered rankings. $\mathcal{S}_{\text{param}}$ specifies optimal dimensionality controls and additional tunable parameters for all 75 descriptor–object-type combinations, serving as reference configurations for the LLM during Reasoning and as deterministic fallback for Action (see Supplementary for full details).

$\mathcal{S}$ is constructed using classification accuracy as the evaluation metric, a standard protocol for topological descriptors [5]. We evaluate on 26 TopoBenchmark datasets (§3.4) with 6 classifiers drawn from the top of the TabArena leaderboard [17] (TabPFN [26], XGBoost [13], CatBoost [47], Random Forest [8], TabM [21], RealMLP [27]) and 5-fold cross-validation. The construction proceeds in three stages. (A) Grid search identifies optimal parameter and dimensionality configurations for each descriptor on each dataset, yielding $\mathcal{S}_{\text{param}}$. (B) Using these optimized configurations, we evaluate all 15 descriptors across 26 datasets and 6 classifiers ($26 \times 15 \times 6 = 2{,}340$ balanced accuracy values), and derive per-object-type tiered rankings from each descriptor's best-performing classifier per dataset, averaged across datasets of the same object type, yielding $\mathcal{S}_{\text{rank}}$. (C) Five domain experts analyzed the quantitative results and formulated the qualitative components of both $\mathcal{S}_{\text{prop}}$ and $\mathcal{S}_{\text{rank}}$: descriptor properties, parameter reasoning heuristics, reasoning chains, and PH signal definitions, all grounded in TDA principles. To prevent information leakage and improve generalization, all components of $\mathcal{S}$ are organized at the object-type level rather than encoding dataset-specific information. The agent never sees dataset identities, class labels, or per-dataset accuracy values at inference time. For example, $\mathcal{S}_{\text{rank}}$ records the persistence landscape ranks in Tier 1 for *glands/lumens*, not that it attains a specific accuracy on any dataset. While $\mathcal{S}$ shares similarities with retrieval-augmented generation (RAG), two key differences distinguish it: asymmetric information access during proposal prevents anchoring (§3.2), and Reflection validates outputs against quality criteria, which standard RAG pipelines lack.

Dual Memory. The static skill set cannot anticipate every possible scenario: unseen imaging modalities, atypical PH profiles, or edge cases may require runtime adaptation. TopoAgent addresses this through dual memory, inspired by the verbal memory mechanism in Reflexion [49]. Short-term memoryM_s records tool invocations within a single run, implemented as the accumulated history within the LLM's context window, preventing repeated failures on the same image. Long-term memory M_l accumulates structured diagnostic entries across runs within a dataset: each entry records the failed descriptor, diagnosed cause, and successful correction, enabling the agent to learn from earlier mistakes on the same dataset. M_l is reset between datasets to prevent cross-dataset leakage but accumulates within each dataset, enabling within-dataset adaptation.

Algorithm 1 summarizes the complete TopoAgent workflow, showing how the pipeline (§3.2) and components (§3.3) interact.

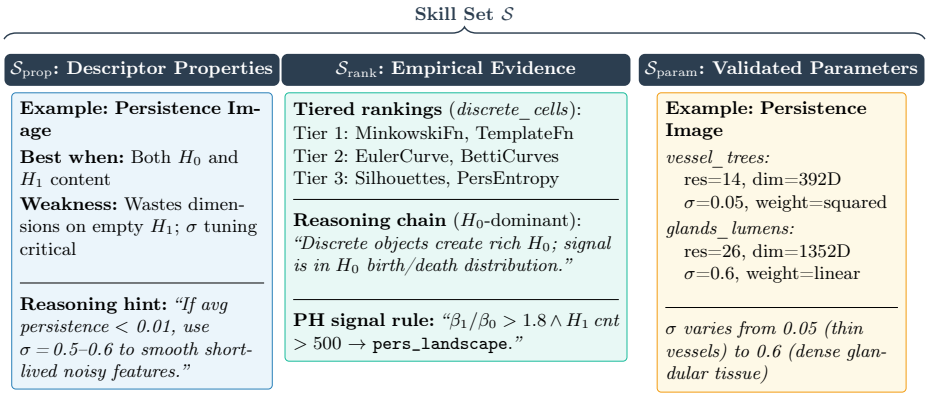

Fig. 3: Examples of the skill set $\mathcal{S}$ using persistence image. $\mathcal{S}_{\text{prop}}$ encodes qualitative descriptor knowledge, $\mathcal{S}_{\text{rank}}$ provides tiered rankings and reasoning chains, and $\mathcal{S}_{\text{param}}$ specifies validated parameters per object type.

3.4 TopoBenchmark

TopoBenchmark serves two purposes: it provides an empirical basis from which the skill set $\mathcal{S}$ is distilled, and it defines a frozen testbed for evaluating the agent's descriptor determinations. As discussed in §1, existing studies lack consistent protocols for cross-descriptor comparison. TopoBenchmark addresses this gap with standardized evaluation across diverse morphologies.

We curate 26 publicly available 2D medical image classification datasets spanning 11 imaging modalities and 11 clinical domains (full details in the Supplementary), organized into five object types that cover the major morphological categories in medical imaging: *cells* (6 datasets), *glands/lumens* (6), *organ shapes* (8), *vessel trees* (3), and *surface lesions* (3). These five categories exhibit distinct topological signatures that motivate adaptive descriptor determination: discrete cells have the most balanced H_0/H_1 ratio ($\beta_1/\beta_0 \approx 1.3$) with the lowest total feature count (∼2,090 persistence pairs per image), vessel trees show the strongest H_1 dominance ($\beta_1/\beta_0 \approx 1.9$) from branching vasculature, and glands/lumens give the highest total feature count (∼4,842) from complex tissue architecture. Organ shapes exhibit the highest within-type variance, from 602 total pairs (OrganAMNIST) to 7,803 (OCTMNIST). No single descriptor performs best across all five types (Table 1), showing that descriptor determination must be adaptive.

The complete collection contains over 1.2 million images, making full evaluation practically prohibitive. Moreover, the dataset sizes vary by three orders of magnitude (516 to 327,680), making a uniform sample size inappropriate: too small wastes discriminative power on large datasets, and too large exceeds the smallest datasets entirely. We address this with convergence analysis: for each dataset, we evaluate the top-3 descriptors at increasing sample sizes $n \in \{50, 100, 200, \ldots, 10{,}000\}$ ($n_{\max}$ capped at the dataset size) using 5-fold cross-validation with TabPFN [26] across 3 seeds, and define n^* as the small-

Algorithm 1 TopoAgent

Require: Medical image I, task prompt, time limit $t_{\max}$
Ensure: Topological feature vector $\mathbf{f} \in \mathbb{R}^{n_d}$, reasoning trace $\mathcal{R}$

$\mathcal{S} \leftarrow$ LoadSkillSet( ); *Tools* $\leftarrow$ {perception tools (6), descriptor tools (15)}
$M_s \leftarrow [\,]$; $M_l \leftarrow$ LoadLongTermMemory( )
— **Perception** —
$(h(I),\ v(I)) \leftarrow$ PerceptionTools(I); $M_s \leftarrow M_s \cup \{h(I),\ v(I)\}$
$o \leftarrow$ LLM.Perceive$(I,\ M_s)$ ▷ Object type $o \in \mathcal{O}$ via vision-enabled LLM
— **Reasoning (Proposal)** —
$(d_p,\ \theta_p) \leftarrow$ LLM.Propose$(h(I),\ v(I),\ o,\ \mathcal{S}_{\text{prop}})$ ▷ Stripped $\mathcal{S}_{\text{rank}}$ only
— **Reasoning (Determination) + Action + Reflection** —
while Elapsed( ) $< t_{\max}$ **do**
  $(d^*,\ \theta^*) \leftarrow$ LLM.Determine$(d_p,\ \theta_p,\ \mathcal{S}_{\text{rank}},\ M_l)$
  $\mathbf{f} \leftarrow$ ExecuteDescriptorTool$(d^*,\ \theta^*)$ ▷ Action: run descriptor tool
  $q \leftarrow$ LLM.Reflect$(\mathbf{f},\ d^*,\ \mathcal{S},\ M_l)$ ▷ Quality diagnosis
  if $q.pass$ **then return** $(\mathbf{f},\ \{h(I), v(I), o, d_p, d^*, \theta^*, q\})$
  end if
  $M_l.update(d^*,\ q)$ ▷ Record diagnosis + correction
end while
$(\mathbf{f},\ \mathcal{R}) \leftarrow$ Fallback$(\mathcal{S}_{\text{rank}},\ o)$ ▷ Top-ranked descriptor for o
return $(\mathbf{f},\ \mathcal{R})$

est n satisfying three criteria simultaneously: balanced accuracy within 1% of the value at $n_{\max}$, standard deviation across the seeds below 2%, and descriptor ranking agreement (Spearman $\rho = 1.0$ among the top-3 descriptors). For example, PCam (5,000 histopathology patches) requires the full $n^* = 5{,}000$ for accuracy convergence, while IDRiD (516 retinal images) converges at $n^* = 500$. However, descriptor ranking stabilizes much earlier ($n = 100$ for PCam, and $n = 50$ for IDRiD), meaning the relative ordering of the descriptors converges well before the absolute accuracy plateaus. The resulting frozen benchmark contains 113,182 samples with per-dataset sizes of 500–7,500, pre-computed persistent homology caches, and fixed fold indices for reproducible evaluation (the construction pipeline in the Supplementary).

4 Experiments

Setup. We evaluate TopoAgent on all 26 TopoBenchmark datasets grouped by the five object types. For each image, the agent receives only the raw image and a task prompt; no dataset identity or object-type label is provided. TopoAgent is implemented using LangGraph [32] with GPT-4o [28] as its backbone LLM. We compare with three categories of baselines: (1) general-purpose LLMs (GPT-4o, Gemini 2.5 Pro [20], Claude Sonnet 4.6 [6]), each equipped with the same 21 topology tools in a ReAct-style loop but without the skill set $\mathcal{S}$, dual memory, and the structured PRAR pipeline; (2) a medical imaging agent (MedRAX [18]) that orchestrates radiological tools but lacks topological capa-

Table 1: Balanced accuracy (%) on TopoBenchmark across the five object types. Bold: the best; underline: the second best.

Method	TopoBenchmark					Avg.
	Cells	Glands	Organs	Lesions	Vessels	
GPT-4o [28]	66.46	44.33	52.32	34.59	33.53	46.25
Gemini 2.5 Pro [20]	63.57	52.19	52.10	33.39	33.53	46.96
Claude Sonnet 4.6 [6]	60.58	51.15	54.47	33.13	33.38	46.54
MedRAX [18]	66.17	62.99	52.66	38.71	32.15	50.54
Persistence image [2]	59.17	47.22	44.03	35.04	36.14	44.32
Persistence statistics [5]	64.43	54.34	52.73	42.76	34.76	49.80
Object-type oracle	<u>73.21</u>	<u>72.16</u>	<u>58.46</u>	<u>50.13</u>	<u>40.50</u>	<u>58.89</u>
TopoAgent (ours)	**79.34**	**81.63**	**68.99**	**59.81**	**51.30**	**68.21**

bilities[3]; (3) fixed-descriptor baselines: persistence image [2], persistence statistics [5], and an object-type oracle that selects the top-ranked descriptor from $\mathcal{S}_{\text{rank}}$ with validated parameters from $\mathcal{S}_{\text{param}}$ per object type, without per-image LLM reasoning. All LLM calls use greedy decoding (temperature 0) for reproducibility. All evaluations use 5-fold cross-validation on frozen splits across six classifiers evaluated during skill set construction. We report the best accuracy among them. TabPFN [26] ranks the highest on the majority of the datasets. Implementation details are in the Supplementary.

Main Results. Table 1 summarizes the balanced accuracy across the five object types. TopoAgent achieves 68.21% average balanced accuracy, outperforming the strongest baseline (object-type oracle, 58.89%) by 9.32%. Since TopoAgent and the GPT-4o baseline share the same underlying LLM and toolset, the 21.96% gap (46.25% vs. 68.21%) highlights the combined contribution of the structured PRAR pipeline, skill set, and dual memory. The three general-purpose LLMs cluster within a narrow range (46.25–46.96%), suggesting that without structured reasoning, LLM capability alone is insufficient. MedRAX [18] improves over the general-purpose LLMs by ∼4% with medical knowledge but lacks topological features and falls nearly 18% below TopoAgent. TopoAgent significantly outperforms all baselines under the Wilcoxon signed-rank test (oracle: $p < 0.01$; all others: $p < 0.001$). Among the fixed-descriptor baselines, the object-type oracle (58.89%) represents the ceiling of non-adaptive determination; the 9.32% gap to TopoAgent demonstrates the value of per-image descriptor determination, with the largest gains on vessels (+10.80%) and organs (+10.53%), where descriptor choice is most sensitive to image-specific PH characteristics.

Ablation Study. Table 2 examines the contribution of each TopoAgent component. Reasoning and Action are not ablated as they are essential for producing any output. Among the four removable components, the skill set $\mathcal{S}$ is the most impactful (8.65% drop); without it, the agent is only marginally better than

[3] EndoAgent [53] was also considered but excluded as it was withdrawn by its authors.

Table 2: Ablation study. Top: component ablation decomposing Perception, Reasoning, skill set, and dual memory. Middle: design ablation. Bottom: backbone ablation.

Variant	TopoBenchmark					Avg.
	Cells	Glands	Organs	Lesions	Vessels	
w/o Perception	76.67	78.62	63.19	44.90	42.74	61.22
w/o Reflection	75.18	80.62	60.15	50.90	41.55	61.68
w/o Skill set $\mathcal{S}$	71.68	79.79	57.81	51.57	36.96	59.56
w/o Dual memory	79.22	81.31	65.58	45.05	43.34	62.90
w/o PH profile $h(I)$	77.59	81.53	66.49	48.38	46.61	64.12
w/o Visual statistics $v(I)$	77.03	80.96	68.13	50.26	48.39	64.95
w/o Proposal step	73.69	73.21	58.04	52.11	41.75	59.76
w/o $\mathcal{S}_{\text{prop}}$	76.83	80.08	63.93	56.93	42.77	64.11
w/o $\mathcal{S}_{\text{rank}}$	72.52	79.82	59.26	53.85	39.19	60.93
w/o $\mathcal{S}_{\text{param}}$	74.21	79.99	62.41	55.29	41.25	62.63
w/o Short-term M_s	79.25	81.42	67.81	55.35	49.82	66.73
w/o Long-term M_l	79.26	81.36	66.42	48.89	46.34	64.45
TopoAgent (Claude 4.6)	78.15	**82.17**	**70.42**	57.29	**53.82**	**68.37**
TopoAgent (Gemini 2.5 Pro)	76.93	80.95	70.25	59.14	52.65	67.98
TopoAgent (Llama3 8B [22])	70.98	76.63	56.41	41.63	40.89	57.31
TopoAgent (GPT-4o)	**79.34**	81.63	68.99	**59.81**	51.30	68.21

object-type oracle (59.56% vs. 58.89%), confirming that $\mathcal{S}$ provides an essential knowledge base. Perception (6.99% drop) is critical for lesions and vessels, where object-type identification drives descriptor choice. Reflection (6.53% drop) matters the most for organs and lesions, which quite often require parameter correction. Dual memory shows the smallest component-level drop (5.31%) but is critical for lesions (14.76%) and vessels (7.96%), whose PH profiles deviate the most from the skill set's general rankings. The design ablation further decomposes each component. Among the skill set components, $\mathcal{S}_{\text{rank}}$ individually has the largest impact (7.28% drop), followed by $\mathcal{S}_{\text{param}}$ (5.58%) and $\mathcal{S}_{\text{prop}}$ (4.10%), confirming that empirical rankings carry more weight than descriptive properties. Removing the Proposal step drops the performance to 59.76%, very close to object-type oracle (58.89%), validating the anti-anchoring design: without an independent proposal, the agent reduces to following $\mathcal{S}_{\text{rank}}$ directly, losing the per-image adaptation ability that accounts for the 9.32% gap. The backbone ablation shows a narrow spread across the models (67.98–68.37%), reflecting the design intent: $\mathcal{S}$ encodes domain expertise that any frontier LLM can leverage. However, Llama3 8B [22] drops to 57.31%, below object-type oracle, indicating that frontier-level reasoning is necessary to benefit from the pipeline. The agent's value lies in the PRAR pipeline that orchestrates tool use, skill retrieval, and quality validation, which zero-shot LLMs lack (trailing by over 21%).

Case Study. In Fig. 4, two *organ shape* images share the same object type and the same $\mathcal{S}_{\text{rank}}$ recommendation (`minkowski_functionals`, Tier 1), yet the

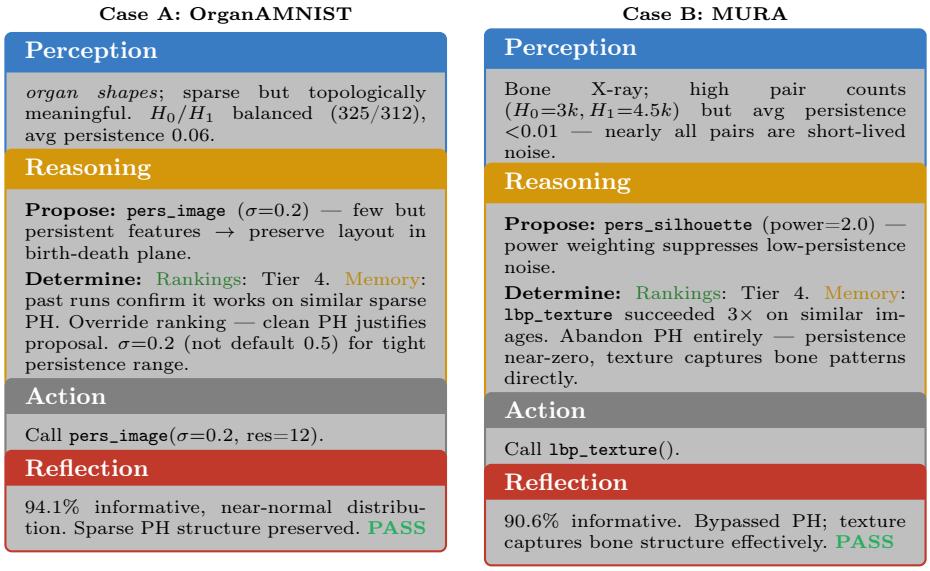

Fig. 4: Two case studies on *organ shapes* with opposite PH profiles. **Case A:** sparse but meaningful PH (avg persistence=0.06) — the agent retains a Tier 4 proposal with data-driven parameters. **Case B:** noisy PH (avg persistence<0.01) — the agent abandons PH-based descriptors entirely. Same object type, same ranking, opposite determinations driven by image-specific PH signals.

agent arrives at different determinations. In Case A (avg persistence = 0.06), the agent retains a Tier 4 proposal based on sparse but meaningful PH and memory confirmation. In Case B (avg persistence < 0.01), the agent abandons PH-based descriptors entirely in favor of texture, guided by M_l. A rule-based system would assign the same descriptor to both images based on object type alone. TopoAgent distinguishes them: Proposal reads the PH profile, Determination weighs it against rankings and memory, and Reflection verifies the outcome.

Downstream Integration. Table 3 evaluates whether TopoAgent's determinations improve DL pipelines by integrating the chosen descriptors into ResNet-152 [25] and SwinV2-B [38] following the PHG-Net [45] fusion protocol on ISIC 2018 [64], Prostate Cancer [33], and CBIS-DDSM [34]. The fixed persistence image consistently *hurts* performance relative to the backbone alone, confirming that a mismatched descriptor can be worse than no topology at all. TopoAgent matches or exceeds PHG-Net on 4 of 6 backbone/dataset combinations and remains within one standard deviation on the other two, using pre-computed topological feature vectors that require no differentiable PH layer, while the determined Minkowski functionals provide interpretable geometric signatures (area, perimeter, Euler characteristic) that PHG-Net's learned features lack. Although TopoAgent determines Minkowski functionals for all three external datasets, the parameters differ per dataset, reflecting $\mathcal{S}_{\text{prop}}$-guided reasoning on unseen data rather than $\mathcal{S}_{\text{rank}}$ lookup, as no benchmark rankings exist for these datasets. All

Table 3: Downstream integration with CNN and Transformer backbones. PI = fixed persistence image; PHG = PHG-Net's [45] learnable topological features; TopoAgent = adaptively determined descriptor. Accuracy (%) is reported as mean±std with 5 runs. TopoAgent determines Minkowski functionals for all three datasets.

Method	ISIC	Prostate	CBIS-DDSM
ResNet-152 [25]	88.76±0.37	93.14±0.33	72.01±0.51
ResNet-152 + PI	87.91±0.11	93.05±0.12	70.45±0.18
ResNet-152 + PHG	89.98±0.21	93.98±0.31	**74.51±0.29**
ResNet-152 + TopoAgent	**90.13±0.13**	**94.22±0.18**	74.26±0.14
SwinV2-B [38]	90.57±0.29	95.87±0.25	73.79±0.31
SwinV2-B + PI	90.45±0.13	95.86±0.16	73.13±0.18
SwinV2-B + PHG	91.88±0.18	**98.55±0.31**	77.08±0.30
SwinV2-B + TopoAgent	**92.01±0.23**	97.89±0.24	**77.13±0.24**

three datasets are external to TopoBenchmark and were not used during skill set construction, confirming that the agent generalizes beyond its training data.

Efficiency. For 86.9% of images, the PRAR pipeline completes in a single pass with 4 LLM calls and 29.0 s wall-clock time. Reflection triggers a retry in the remaining 13.1%, adding on average 1.8 calls and 22.1 s. Overall, TopoAgent requires 4.2 LLM calls and 31.9 s per image on average, making it practical for batch processing across large datasets.

5 Conclusions

We presented TopoAgent, the first LLM-based agentic framework for automated topology learning in medical imaging. By combining a Perception–Reasoning–Action–Reflection loop with a distilled skill set and dual memory, TopoAgent addresses a critical bottleneck: determining effective descriptors currently requires extensive domain expertise and trial-and-error experimentation. On TopoBenchmark (26 datasets, five object types), TopoAgent achieves 68.21% average balanced accuracy, outperforming the strongest baseline by 9.32% while providing interpretable topological feature vectors. Downstream integration with both CNN and Transformer backbones confirms that its pre-computed features match or exceed learnable topological representations without additional training.

As the skill set is distilled from 26 2D medical image datasets, it may not generalize to fundamentally different topological structures (e.g., 3D volumetric data or non-medical domains). The agent also inherits the latency and cost of LLM inference (∼32s per image). Future work includes extending TopoBenchmark to 3D images, incorporating the agent into end-to-end training pipelines, and exploring multi-descriptor fusion guided by the agent's reasoning.

Acknowledgments. The authors gratefully acknowledge support from the National Science Foundation under Grants CCF-2444309 and CCF-2523787, and from the American Heart Association under Award 26AIREA1574568.

References

1. Adame, D., Nunez, J.A., Vazquez, F., Gurrola, N., Li, H., Tang, H., Fu, B., Gu, P.: Topo-VM-UNetV2: Encoding topology into vision Mamba UNet for polyp segmentation. In: CBMS. pp. 258–263 (2025)
2. Adams, H., Emerson, T., Kirby, M., Neville, R., Peterson, C., et al.: Persistence Images: A stable vector representation of persistent homology. Journal of Machine Learning Research **18**(8), 1–35 (2017)
3. Ahmed, F.: Topological signatures vs. gradient histograms: A comparative study for medical image classification. arXiv preprint arXiv:2507.03006 (2025)
4. Ahmed, F., Bhuiyan, M.A.N., Coskunuzer, B.: Topo-CNN: Retinal image analysis with topological deep learning. Journal of Imaging Informatics in Medicine (2025)
5. Ali, D., Asaad, A., Jimenez, M.J., Nanda, V., Paluzo-Hidalgo, E., Soriano-Trigueros, M.: A survey of vectorization methods in topological data analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(12) (2023)
6. Anthropic: The Claude model card (2025), `https://www.anthropic.com/research/claude-model-card`
7. Asia Pacific Tele-Ophthalmology Society: APTOS 2019 blindness detection (2019), `https://www.kaggle.com/c/aptos2019-blindness-detection`
8. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)
9. Bubenik, P.: Statistical topological data analysis using persistence landscapes. The Journal of Machine Learning Research **16**(1), 77–102 (2015)
10. Carlsson, G.: Topology and data. Bulletin of the American Mathematical Society **46**(2), 255–308 (2009)
11. Carrière, M., Chazal, F., Ike, Y., Lacombe, T., Royer, M., Umeda, Y.: PersLay: A neural network layer for persistence diagrams and new graph topological signatures. In: AISTATS. pp. 2786–2796 (2020)
12. Chambers, E.W., Meng, G.: A stable and theoretically grounded Gromov-Wasserstein distance for Reeb graph comparison using persistence images. arXiv preprint arXiv:2507.01171 (2025)
13. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: ACM SIGKDD. pp. 785–794 (2016)
14. Cheng, J., Huang, W., Cao, S., Yang, R., Yang, W., Yun, Z., Wang, Z., Feng, Q.: Enhanced performance of brain tumor classification via tumor region augmentation and partition. PLoS ONE **10**(10), e0140381 (2015)
15. Combalia, M., Codella, N.C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S., Malvehy, J.: Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: The 2019 International Skin Imaging Collaboration Grand Challenge. The Lancet Digital Health **4**(5), e330–e339 (2022)
16. Edelsbrunner, Letscher, Zomorodian: Topological persistence and simplification. Discrete & Computational Geometry **28**(4), 511–533 (2002)
17. Erickson, N., Purucker, L., Tschalzev, A., Holzmüller, D., Desai, P.M., Salinas, D., Hutter, F.: TabArena: A living benchmark for machine learning on tabular data. arXiv preprint arXiv:2506.16791 (2025)
18. Fallahpour, A., Ma, J., Munim, A., Lyu, H., Wang, B.: MedRAX: Medical reasoning agent for chest X-Ray. arXiv preprint arXiv:2502.02673 (2025)
19. Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., Hutter, F.: Efficient and robust automated machine learning. Advances in Neural Information Processing Systems **28** (2015)

20. Google DeepMind: Gemini 2.5: Our most intelligent AI model (2025), https://deepmind.google/technologies/gemini/
21. Gorishniy, Y., Kotelnikov, A., Babenko, A.: TabM: Advancing tabular deep learning with parameter-efficient ensembling. In: ICLR (2025)
22. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The Llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
23. Gu, P., Li, H., Tang, H., Xu, D., Enriquez, E., Kim, D., Fu, B., Chen, D.Z.: Integrating multi-scale and multi-filtration topological features for medical image classification. In: WACV. pp. 8660–8669 (2026)
24. Gu, P., Wang, H., Zhang, Y., Li, H., Wang, C., Chen, D.: TopoImages: Incorporating local topology encoding into deep learning models for medical image classification. In: ACM MM. pp. 1938–1947 (2025)
25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
26. Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., et al.: Accurate predictions on small data with a tabular foundation model. Nature (2025)
27. Holzmüller, D., Grinsztajn, L., Obst, N.: Better by default: Strong pre-tuned MLPs and boosted trees on tabular data. In: NeurIPS (2024)
28. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., et al.: GPT-4o system card. arXiv preprint arXiv:2410.21276 (2024)
29. Hutter, F., Kotthoff, L., Vanschoren, J.: Automated Machine Learning: Methods, Systems, Challenges. Springer (2019)
30. Khamis, A.: Agentic AI systems: Architecture and evaluation using a frictionless parking scenario. IEEE Access (2025)
31. Kim, K., Kim, J., Zaheer, M., Kim, J., Chazal, F., Wasserman, L.: PLLay: Efficient topological layer based on persistent landscapes. Advances in Neural Information Processing Systems **33**, 15965–15977 (2020)
32. LangChain, Inc.: LangGraph: Build resilient language agents as graphs (2024), https://github.com/langchain-ai/langgraph
33. Lawson, P., Sholl, A.B., Brown, J.Q., Fasy, B.T., Wenk, C.: Persistent homology for the quantitative evaluation of architectural features in prostate cancer histology. Scientific Reports **9**(1), 1139 (2019)
34. Lee, R.S., Gimenez, F., Hoogi, A., Miyake, K.K., Gorovoy, M., Rubin, D.L.: A curated mammography data set for use in computer-aided detection and diagnosis research. Scientific Data **4**(1), 1–9 (2017)
35. Li, B., Yan, T., Pan, Y., Luo, J., Ji, R., et al.: MMedAgent: Learning to use medical tools with multi-modal agent. In: Findings of EMNLP. pp. 8745–8760 (2024)
36. Liang, P., Ding, Y., Zhang, Y., Chen, J., Zheng, H., Wang, H., Zhang, Y., Meng, G., Weninger, T., Niemier, M., et al.: Cell instance segmentation: The devil is in the boundaries. IEEE Transactions on Medical Imaging **45**(4), 1311–1324 (2026)
37. Liu, J., Shen, L., Wei, G.W.: ChatGPT for computational topology. Foundations of Data Science **6**(2), 221 (2024)
38. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al.: Swin Transformer V2: Scaling up capacity and resolution. In: CVPR. pp. 12009–12019 (2022)
39. Maria, C., Boissonnat, J.D., Glisse, M., Yvinec, M.: GUDHI: Simplicial complexes and persistent homology packages. In: ICMS. pp. 167–174 (2014)
40. Mecke, K.R.: Additivity, convexity, and beyond: Applications of Minkowski functionals in statistical physics. In: SP-SS, pp. 111–184. Springer (2000)

41. Meng, G., Gu, P., Liang, P., Lalor, J.P., Chambers, E.W., Chen, D.Z.: TopoCL: Topological contrastive learning for medical imaging. In: CVPR. pp. 42681–42690 (2026)
42. Meng, G., Zeng, Q., Lalor, J.P., Yu, H.: A psychology-based unified dynamic framework for curriculum learning. Computational Linguistics pp. 1–49 (2025)
43. Meng, G., Zhou, R., Liu, L., Liang, P., Liu, F., Chen, D.Z., Niemier, M., Hu, X.S.: Efficient approximation of Earth Mover's distance based on nearest neighbor search. IEEE Transactions on Multimedia (2025)
44. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(7), 971–987 (2002)
45. Peng, Y., Wang, H., Sonka, M., Chen, D.Z.: PHG-Net: Persistent homology guided medical image classification. In: WACV. pp. 7583–7592 (2024)
46. Perea, J.A., Munch, E., Khasawneh, F.A.: Approximating continuous functions on persistence diagrams using template functions. Foundations of Computational Mathematics **23**(4), 1215–1272 (2023)
47. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: CatBoost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems **31** (2018)
48. Royer, M., Chazal, F., Levrard, C., Umeda, Y., Ike, Y.: ATOL: Measure vectorization for automatic topologically-oriented learning. In: AISTATS (2021)
49. Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., Yao, S.: Reflexion: Language agents with verbal reinforcement learning. Advances in Neural Information Processing Systems **36**, 8634–8652 (2023)
50. Singhal, S., Li, C., Aukerman, A., Carrière, M., Miller, M.L., Hibshoosh, H., McDonald, J.A., Winfield, J.R., et al.: Topology-based biomarkers accurately predict breast cancer outcome and survival. Cancer Research (2026)
51. Spanhol, F.A., Oliveira, L.S., Petitjean, C., Heutte, L.: A dataset for breast cancer histopathological image classification. IEEE Transactions on Biomedical Engineering **63**(7), 1455–1462 (2016)
52. Su, Z., Liu, X., Hamdan, L.B., Maroulas, V., Wu, J., Carlsson, G., Wei, G.W.: Topological data analysis and topological deep learning beyond persistent homology: A review. Artificial Intelligence Review (2025)
53. Tang, Y., Wang, K., Chen, Y., Zhou, G.: EndoAgent: A memory-guided reflective agent for intelligent endoscopic vision-to-decision reasoning. arXiv preprint arXiv:2508.07292 (2025), withdrawal
54. Turner, K., Mukherjee, S., Boyer, D.M.: Persistent homology transform for modeling shapes and surfaces. Information and Inference: A Journal of the IMA **3**(4), 310–344 (2014)
55. Umeda, Y.: Time series classification via topological data analysis. Information and Media Technologies **12**, 228–239 (2017)
56. Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., thers: Voyager: An open-ended embodied agent with large language models. arXiv preprint arXiv:2305.16291 (2023)
57. Wang, H., Huang, G., Zhao, Z., Cheng, L., et al.: CCF-GNN: A unified model aggregating appearance, microenvironment, and topology for pathology image classification. IEEE Transactions on Medical Imaging (2023)
58. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., et al.: A survey on large language model based autonomous agents. Frontiers of Computer Science **18**(6), 186345 (2024)

59. Wang, Z., Wu, J., Cai, L., Low, C.H., Yang, X., Li, Q., Jin, Y.: MedAgent-Pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. arXiv preprint arXiv:2503.18968 (2025)
60. Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., et al.: The rise and potential of large language model based agents: A survey. Science China Information Sciences **68**(2), 121101 (2025)
61. Xu, G., Li, X., Chen, Y., Duan, Y., Wu, S., Yu, A., Chiu, C.H., et al.: A comprehensive survey of agentic AI in healthcare. Authorea Preprints (2025)
62. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: MedMNIST v2–a large-scale lightweight benchmark for 2D and 3D biomedical image classification. Scientific Data **10**(1), 41 (2023)
63. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K.R., Cao, Y.: ReAct: Synergizing reasoning and acting in language models. In: ICLR (2022)
64. Zhuang, J., Li, W., Manivannan, S., Wang, R., et al.: Skin lesion analysis towards melanoma detection using deep neural network ensemble. ISIC Challenge (2018)
65. Zia, A., Khamis, A., Nichols, J., Tayab, U.B., Hayder, Z., Rolland, V., Stone, E., Petersson, L.: Topological deep learning: A review of an emerging paradigm. Artificial Intelligence Review **57**(4), 77 (2024)