BAYESIAN RIPS ACTIVE LEARNING: TOPOLOGY-AWARE ACQUISITION FOR RARE LINEAGES

Weixiao Wang
Department of Mathematics
Western University
London, ON, Canada
wwang987@uwo.ca

Anibal M. Medina-Mardones
Department of Mathematics
Western University
London, ON, Canada
anibal.medina.mardones@uwo.ca

ABSTRACT

Rare cellular lineages form disconnected or weakly-connected components in gene-expression space, yet standard active learning (AL) acquisitions concentrate queries in dense regions and fail to sample across them. We propose **Bayesian Rips Active Learning (BRAL)**, which augments Gaussian process predictive entropy with a topological signal: each candidate is scored by how much it would change the Vietoris–Rips persistent H_0 of the currently labeled set, favoring selections that expand into disconnected regions. On Paul15 hematopoietic data ($\sim$2.5% rare Megakaryocytes), BRAL discovers 42 of 68 rare cells at a 5% budget—nearly twice the best baseline. On PBMC 3k, BRAL achieves the highest rare-class F1 (0.89) with the lowest variance across seeds.

Single-cell RNA sequencing enables cell-type discovery at unprecedented resolution (Regev et al., 2017; Luecken & Theis, 2019), but rare populations—fate-committed progenitors, disease-associated subpopulations—remain hard to find and expensive to label (Grün et al., 2015; Paul et al., 2015). Active learning (AL) reduces annotation cost by selecting the most informative cells for expert labeling (Settles, 2009), yet recent single-cell benchmarks show that standard AL methods fail to discover rare populations under realistic budgets (Geuenich et al., 2024).

The failure is structural. Rare lineages form disconnected or weakly-connected components in expression space—thin branches or isolated clusters separated by density gaps. Uncertainty sampling (Houlsby et al., 2011; MacKay, 1992) queries ambiguous points near existing decision boundaries; geometric coverage (Sener & Savarese, 2018) maximizes distances but is blind to connectivity; gradient-based diversity (Ash et al., 2020) operates in parameter space without tracking manifold structure. None explicitly favor sampling patterns that reach into topologically distinct regions (extended related work in Appendix A).

We propose to augment acquisition with a *topological signal* derived from the labeled set itself: if labeling a candidate substantially changes the connected-component structure (H_0) of what has been labeled so far, that point likely lies in an unexplored region. This insight leads to **Bayesian Rips Active Learning (BRAL)**, which combines GP entropy with a Vietoris–Rips H_0 bottleneck-distance term.

We make three contributions:

1. We introduce the use of the persistent topology of the labeled set as an acquisition signal that encourages exploration of topologically distinct regions of the data manifold.

2. We realize this idea in a concrete algorithm, Bayesian Rips Active Learning (BRAL), which combines GP entropy with Vietoris–Rips persistent H_0.

3. We evaluate BRAL on synthetic data and two real scRNA-seq datasets, demonstrating that standard baselines (including region-based topological AL) exhibit cold-start failure on rare lineages, while BRAL achieves the most consistent performance across different topologies.

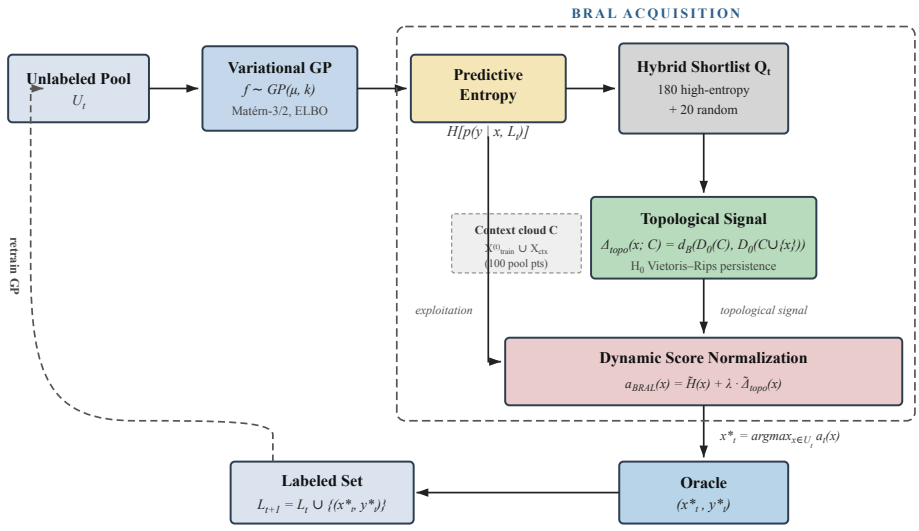

Figure 1: BRAL acquisition pipeline. At each query step, the GP classifier produces predictive entropy over the pool. A hybrid shortlist (180 high-entropy + 20 random) is scored by topological impact Δ_{topo}—the change in H_0 Rips persistence of a context cloud $C = X^{(t)}_{\text{train}} \cup X_{\text{ctx}}$ upon adding each candidate, where X_{ctx} is a random subsample of 100 pool points providing manifold structure. Dynamic Score Normalization combines both terms into the final acquisition score. The topology of accumulated labels drives the next selection.

1 METHOD

1.1 PROBLEM SETUP

We consider pool-based binary AL with an imbalanced rare class. Let $U = \{x_i\}_{i=1}^N$ be the unlabeled pool and $L_t = \{(x_j, y_j)\}$ the initial labeled set, with $y_j \in \{0, 1\}$. At each step t, the acquisition function selects $x^\star_t = \arg\max_{x \in U_t} a_t(x)$, after which $(x^\star_t, y^\star_t)$ is added to L_t.

1.2 VARIATIONAL GP CLASSIFIER

We model the labeling function with a GP prior and Bernoulli likelihood: $f \sim \mathcal{GP}(\mu, k)$, where k is a Matérn-$\frac{3}{2}$ kernel. Since exact inference is intractable, we use variational inference(Hensman et al., 2014) via GPyTorch (Gardner et al., 2018): a Cholesky variational distribution with all training points as inducing points, optimized by maximizing the ELBO. The model is retrained from scratch after each query.

1.3 GUIDING SELECTION VIA THE TOPOLOGY OF LABELED POINTS

The key idea behind BRAL is to let the evolving topology of the labeled set itself guide which point to query next. At iteration t, let $S = X^{(t)}_{\text{train}}$ denote the current labeled features. Its H_0 Vietoris–Rips persistence diagram $D_0(S)$ summarizes the multi-scale connected component structure of what has been labeled so far: how many clusters exist, at what scale they merge, and where topological gaps remain (Appendix N).

For each candidate x, we measure how much adding it would change this topology. Because computing Rips persistence on S alone is unreliable when few labels exist (the complex is too sparse to reflect manifold structure), we augment S with a random subsample of 100 pool points to form a context cloud $C = S \cup X_{\text{ctx}}$, which provides ambient manifold scaffolding. We then define the topological impact:

$$\Delta_{\text{topo}}(x;\, C) = d_B\big(D_0(C),\; D_0(C \cup \{x\})\big),$$

where d_B is the bottleneck distance between persistence diagrams, stable under small perturbations (Cohen-Steiner et al., 2007). A large Δ_{topo} indicates that x lies in a region topologically distinct from the current context—it would create a new connected component, bridge a gap between existing clusters, or extend into a sparse branch. In contrast, a candidate near already-labeled points barely changes D_0 and receives a low score. This mechanism explicitly favors sampling patterns that spread across disconnected regions of the data manifold, precisely the behavior needed to discover rare lineages that form their own topological components.

For computational efficiency, we restrict evaluation to a hybrid shortlist Q_t of 200 candidates: 180 with the highest GP entropy plus 20 drawn uniformly at random to encourage exploration of low-uncertainty regions (Appendix D). The base Rips complex on C is computed once per query; only candidate-augmented complexes are recomputed per candidate.

1.4 BRAL: Acquisition Function

BRAL combines exploitation (uncertainty) with a topological signal (Figure 1).

$$a_{\text{BRAL}}(x) = \underbrace{\widetilde{H}(x)}_{\text{exploitation}} + \lambda \cdot \underbrace{\widetilde{\Delta}_{\text{topo}}(x)}_{\text{topological signal}}. \tag{1}$$

Both terms are min–max normalized on the current shortlist Q_t at each iteration:

$$\widetilde{H}(x) = \frac{H(x) - H_{\min}}{H_{\max} - H_{\min} + \epsilon}, \qquad \widetilde{\Delta}_{\text{topo}}(x) = \frac{\Delta_{\text{topo}}(x) - \Delta_{\min}}{\Delta_{\max} - \Delta_{\min} + \epsilon}. \tag{2}$$

Here $H_{\min}, H_{\max}, \Delta_{\min}, \Delta_{\max}$ are taken over Q_t. This per-iteration rescaling places both scores in $[0, 1]$. We fix $\lambda = 1$ in all experiments (Appendix C).

The entropy term exploits classifier uncertainty to refine the decision boundary. The topological term prioritizes regions structurally absent from the labeled set. Each new label reshapes $D_0(C)$, creating a feedback loop in which the topology of accumulated observations drives subsequent queries.

2 Experiments

Apart from a synthetic dataset consisting of a Y-shaped trajectory (n=600, 20-dim, $\sim$9% rare), we evaluate on two real single-cell datasets that exemplify our two motivating topologies: Paul15 (Paul et al., 2015): mouse hematopoietic progenitors ($n\approx$2,730, 50 PCs), where the rare class—Megakaryocytes ($\sim$ 2.5%)—sits on a thin branch continuously connected to the main myeloid trajectory, as confirmed by graph abstraction (Wolf et al., 2019). *PBMC 3k* (10X Genomics): Dendritic cells ($\sim$37 cells, $\sim$7.2%) versus CD14+ Monocytes ($\sim$480, 50 PCs), where the rare class forms a clearly disconnected cluster.

We compare BRAL to the following baseline methods: Random (uniform sampling), BALD (Houlsby et al., 2011) (GP mutual information), CoreSet (Sener & Savarese, 2018) (max–min distance), BADGE (Ash et al., 2020) (gradient-embedding norm), regionPTR (Hadjadj et al., 2023) (region-based topological AL via persistence clustering (Chazal et al., 2013)), and TopoOnly (BRAL with $\lambda \to \infty$).

The specific protocol for this comparison is as follows. We run 10 random seeds per dataset. The labeling budget is $N \approx 5\%$ of the pool (Paul15: $N = 100$; PBMC and Synthetic: $N = 20$). Cold-start uses 5 labeled points (2 positive, 3 negative). All methods share the same GP oracle, retrained at each acquisition step. We report rare-class F1, cumulative discoveries (Disc), and average precision (AP), with significance assessed via paired Wilcoxon tests. Results are presented in Table 1.

2.1 Key Results

Breaking the cold-start barrier. Table 1 reveals a severe cold-start failure on the branching topology: four of six baselines obtain F1=0.00. BALD discovers 17 rare cells but fails to learn the boundary

	Synthetic (9.0%)			Paul15 (Mk, 2.5%)			PBMC 3k (DC, 7.2%)		
Method	F1	Disc	AP	F1	Disc	AP	F1	Disc	AP
Random	$.67 \pm .27$	4.2	.85	$.00 \pm .00$	4.8	.36	$.00 \pm .00$	3.0	.87
BALD	$.78 \pm .13$	7.3	.93	$.03 \pm .06$	17.0	.75	$.21 \pm .26$	6.7	.99
CoreSet	$.85 \pm .06$	7.3	.93	$.00 \pm .00$	13.6	.50	$.02 \pm .05$	5.8	.94
BADGE	$.69 \pm .22$	10.9	.88	$.31 \pm .16$	22.9	.57	$.62 \pm .41^{\dagger}$	8.3	.95
regionPTR	$.74 \pm .16$	8.9	.89	$.00 \pm .00$	13.5	.60	$.72 \pm .25^{\dagger}$	**13.5**	.98
TopoOnly	$.85 \pm .07$	10.4	.94	$.00 \pm .00$	6.5	.64	$.00 \pm .00$	2.9	.85
BRAL	$\mathbf{.94 \pm .03}$	**11.8**	**.97**	$\mathbf{.45 \pm .07}$	**41.6**	**.76**	$\mathbf{.89 \pm .07}$	12.3	**.99**

Table 1: Comparison of active learning strategies across all datasets. F1 = F1@0.5 (mean $\pm$ std), Disc = rare-class discovery count, AP = average precision. Best results in **bold**. † Not significant vs. BRAL at $p < 0.05$ (Wilcoxon); all unmarked comparisons are significant.

(F1=0.03); BADGE reaches F1=0.31. BRAL attains F1=0.45 and discovers $\sim$42 of 68 rare cells ($p<0.05$ vs. all baselines). The topological term detects a missing component in $D_0(C)$; sampling the Megakaryocyte branch produces a large Δ_{topo} that overrides low entropy.

Disconnected cluster topology. On PBMC 3k, BRAL achieves F1=0.89 ± 0.07 with the lowest variance (vs. 0.41 for BADGE and 0.25 for regionPTR). RegionPTR finds more rare cells (13.5 vs. 12.3) but yields less stable classification because its region partition does not sharpen the decision boundary (Appendix B).

Ablations and additional results. TopoOnly ($\lambda \to \infty$) discovers rare cells but fails on F1; conversely BALD discovers 17 cells on Paul15 yet attains only F1=0.03—both components are needed. On the synthetic bifurcation, BRAL reaches F1=0.94 ± 0.03 vs. 0.85 for the next-best method. Domain-specific baselines achieve high discovery on PBMC 3k but low F1, exposing a discovery–classification gap that BRAL bridges (Appendix G). On PBMC 3k Full ($\sim$1.4% rare), BRAL discovers all reachable rare cells (26.0 ± 0.0); BADGE attains comparable F1 (0.69 vs. 0.64, n.s.; Appendix L). λ is stable in $[0.1, 1.0]$ (Appendix C); BRAL adds $\sim$2.5s/query (Appendix F). BRAL is robust to cold-start conditions, maintaining $F1 \geq 0.88$ even with zero initial positives (Appendix E).

3 Conclusion

We introduced BRAL, an acquisition function that augments GP entropy with a topological signal, VR persistent H_0, explicitly favoring sampling across disconnected components where rare lineages reside. On branching topologies where all six baselines achieve F1$\leq$0.31, BRAL achieves 0.45 and discovers 42 of 68 rare cells; on disconnected clusters (PBMC 3k), BRAL achieves the highest F1 (0.89) with the lowest variance. The core principle—monitoring the topology of what has been labeled to decide what to label next—creates a feedback loop: each oracle label updates the persistent H_0 of the labeled set, and this updated topology directly reshapes the next acquisition step.

Future work and limitations Our current proof-of-concept method is restricted to persistent H_0, single-point acquisitions, and binary classification, and relies on a GP oracle that limits scalability to very large pools. Extending BRAL to batch-mode acquisition—where topological impact is non-additive—higher-dimensional homology (e.g., H_1 loops for under-sampled cycles), multi-class settings, and integration with reference-based cell annotation tools are natural directions for future work.

Acknowledgments

This work was supported by NSERC grants RES000678 and R7444A03.

References

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations (ICLR)*, 2020.

F. Chazal, Leonidas J. Guibas, Steve Oudot, and Primoz Skraba. Persistence-based clustering in riemannian manifolds. *Journal of the ACM*, 60(6):1–43, 2013.

David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37:103–120, 2007.

Seyda Ertekin, Jian Huang, Léon Bottou, and C. Lee Giles. Learning on the border: Active learning in imbalanced data classification. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM)*, pp. 127–136, 2007.

Emanuel Flores-Bautista and Matt Thomson. Unraveling cell differentiation mechanisms through topological exploration of single-cell developmental trajectories. *bioRxiv preprint 2023.07.28.551057*, 2023.

Jacob R Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew Gordon Wilson. GPyTorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Michael J. Geuenich et al. The impacts of active and self-supervised learning on efficient annotation of single-cell expression data. *Nature Communications*, 15:1014, 2024. doi: 10.1038/s41467-024-45198-y.

Dominic Grün, Anna Lyubimova, Lennart Kester, et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, 525(7568):251–255, 2015.

Lies Hadjadj, Emilie Devijver, Remi Molinier, and Massih-Reza Amini. Pool-based active learning with proper topological regions, 2023. URL https://arxiv.org/abs/2310.01597.

James Hensman, Alex Matthews, and Zoubin Ghahramani. Scalable variational gaussian process classification, 2014. URL https://arxiv.org/abs/1411.2005.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. In *Advances in Neural Information Processing Systems*, 2011.

Tram Huynh and Zixuan Cang. Topological and geometric analysis of cell states in single-cell transcriptomic data. *Briefings in Bioinformatics*, 25(3):bbae176, 2024. doi: 10.1093/bib/bbae176.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep Bayesian active learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Suraj Kothawade, Nathan Beck, KrishnaTeja Killamsetty, and Rishabh K. Iyer. SIMILAR: submodular information measures based active learning in realistic scenarios. *CoRR*, abs/2107.00717, 2021. URL https://arxiv.org/abs/2107.00717.

Malte D Luecken and Fabian J Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, 2019.

David J. C. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.

Franziska Paul et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, 163(7):1663–1677, 2015.

Aviv Regev, Sarah A Teichmann, Eric S Lander, et al. The human cell atlas. *eLife*, 6:e27041, 2017.

Ammad H. Rizvi, Pablo G. Camara, Ekaterina K. Kandror, Thomas J. Roberts, Iris Schieren, Tom Maniatis, and Raul Rabadan. Single-cell topological rna-seq analysis reveals insights into cellular differentiation and development. *Nature Biotechnology*, 35(6):551–560, June 2017. doi: 10.1038/nbt.3854.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=H1aIuk-RW.

Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

Tatsuya Shiraishi, Tam Le, Hisashi Kashima, and Makoto Yamada. Topological Bayesian Optimization with Persistence Diagrams. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pp. 1483–1490, 2020. doi: 10.3233/FAIA200255.

F. Alexander Wolf, Fiona K. Hamey, Mireya Plass, et al. Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, 20:59, 2019. doi: 10.1186/s13059-019-1663-x.

Fei Zheng, She Zhang, Christopher Churas, Dexter Pratt, Ivet Bahar, and Trey Ideker. HiDeF: identifying persistent structures in multiscale 'omics data. *Genome Biology*, 22(1):21, 2021. doi: 10.1186/s13059-020-02228-4.

A Related Works

A.1 Bayesian and diversity-based Active Learning

BALD (Houlsby et al., 2011) selects points that maximize mutual information between predictions and model parameters. BatchBALD (Kirsch et al., 2019) extends this principle to batch acquisition.

CoreSet (Sener & Savarese, 2018) formulates selection as a k-center problem, prioritizing geometric coverage in feature space. BADGE (Ash et al., 2020) promotes diversity via gradient embeddings in parameter space while retaining an uncertainty bias.

These approaches primarily reason about uncertainty or geometric spread in high-density regions of the data. They do not explicitly track the connectivity or separation structure of the underlying data manifold. As a result, they can systematically undersample thin or isolated rare lineages when labeling budgets are small (Ertekin et al., 2007; Geuenich et al., 2024).

A.2 Topological data analysis in learning and single-cell biology

In single-cell analysis, scTDA used persistent homology to study differentiation trajectories (Rizvi et al., 2017). PAGA builds a graph abstraction that preserves large-scale topological relations between cell populations (Wolf et al., 2019). These methods are descriptive rather than active: they analyze topology but do not use it to decide which points to label next.

Closest to our setting, region-based topological AL (Hadjadj et al., 2023) partitions the pool via persistence clustering (Chazal et al., 2013) and samples at the region level. Because this approach aggregates points into regions, it can collapse thin continuous branches into the main trajectory, obscuring rare lineages as we study in Appendix B. BRAL differs in two key ways. First, it evaluates topological impact *point-wise* rather than at the level of coarse regions. Second, it measures how each candidate would change the persistent H_0 of the *currently labeled set*, making topology an evolving signal that interacts with the labeling process itself.

Shiraishi et al. (Shiraishi et al., 2020) incorporate persistent homology into Gaussian-process–based Bayesian optimization by defining kernels on persistence diagrams and using them to guide sequential experimentation. Their work shows that topological summaries can meaningfully interact with GP uncertainty in a decision-making loop, rather than serving only as post hoc descriptors. However, their objective is function optimization over structured domains (e.g., graphs or materials), not data acquisition for labeling, and the topology they use summarizes the input space as a whole rather than the evolving topology of labeled observations.

HiDeF (Zheng et al., 2021) uses persistent H_0 to detect cell communities that remain stable across clustering resolutions. By tracking long-lived connected components, it reveals subtle and rare cell populations that are missed by single-scale methods. While HiDeF demonstrates the value of H_0 persistence for rare-cell discovery, it is an *offline descriptive* tool rather than an acquisition strategy: topology is analyzed after data are collected, not used to guide labeling.

Huynh and Cang (Huynh & Cang, 2024) study local Vietoris–Rips H_0 persistence on single-cell graphs to quantify connectivity of cell states along differentiation trajectories. They show that peaks in H_0 persistence correspond to transitional or sparsely connected cell states, which are precisely the kinds of regions that standard AL tends to undersample. However, their analysis is again retrospective; it does not prescribe how to actively select informative cells.

Flores-Bautista and Thomson (Flores-Bautista & Thomson, 2023) develop TopGen, which uses persistent homology (including higher homology) to detect non-tree-like structures such as loops in developmental trajectories. This work illustrates that single-cell manifolds can possess richer topology than simple branching trees, motivating topology-aware methods beyond graph abstractions like PAGA. Nevertheless, TopGen focuses on understanding lineage structure rather than on reducing annotation cost.

Taken together, these works support the premise that persistent homology captures biologically meaningful structure in single-cell data. BRAL differs by making persistence an *active acquisition signal*: instead of merely analyzing topology, it uses changes in persistent H_0 of the labeled set to decide which cells to query next.

B Topological Granularity: Point-wise vs. Region-wise

We implemented **regionPTR**, a baseline inspired by Proper Topological Regions methods (Hadjadj et al., 2023). To ensure a fair comparison, regionPTR operates on fixed topological structures derived from the initial data, avoiding the computational cost of re-clustering at every step.

B.1 Implementation Details

The regionPTR acquisition function selects points in three stages:

1. **Offline Topological Partitioning:** Before the active learning loop begins, we partition the entire unlabeled pool U into disjoint topological regions $R_1, \ldots, R_k$ using the ToMATo algorithm (Chazal et al., 2013) (Topological Mode Analysis Tool) from GUDHI, with `density_type="DTM"`, k=30 neighbors, and merge radius r=0.5. The density is estimated via the distance-to-measure (distance to the k-th nearest neighbor). Regions are fixed throughout the active learning process.
2. **Region Selection (Uncertainty):** At each query step t, we model the proportion of rare-class cells in each region R_j using a Beta-Binomial conjugate prior. Let α_j and β_j be the counts of positive (rare) and negative (common) labels currently observed in region R_j (initialized to 1). We select the target region R^* that maximizes the posterior variance of the rare-class proportion:

$$R^* = \arg\max_{R_j} \frac{\alpha_j \beta_j}{(\alpha_j + \beta_j)^2(\alpha_j + \beta_j + 1)} \tag{3}$$

3. **Point Selection (Geometric Score):** Within the selected region R^*, we select the specific candidate x^* that maximizes a geometric score combining sparsity and peripherality. The DTM proxy is the Euclidean distance to the k-th nearest neighbor. Both terms are min-max normalized:

$$x^* = \arg\max_{x \in R^*} [0.5 \cdot \text{Norm}(\text{DTM}_k(x)) + 0.5 \cdot \text{Norm}(\|x - \bar{x}\|)] \tag{4}$$

where $\text{DTM}_k(x)$ is the distance from x to its k-th nearest neighbor (k=30) and $\bar{x}$ is the pool centroid.

B.2 Results

We compare BRAL and regionPTR on both the continuous branching task (Paul15, N=100) and the discrete cluster task (PBMC 3k subset, N=20). The results are summarized in Table 2.

On Paul15, regionPTR fails ($F_1 = 0.00$) because the unsupervised ToMATo clustering merges the geometrically continuous rare branch into a single large region, preventing region-level uncertainty

Table 2: Comparison of active learning strategies on real datasets. F1 = F1@0.5 (mean ± std), Disc = rare-class discovery count, AP = average precision. Best results in **bold**.

	Paul15 (Mk, 2.5%)			PBMC 3k (DC, 7.2%)		
Method	F1	Disc	AP	F1	Disc	AP
Random	.00 ± .00	4.8	.36	.00 ± .00	3.0	.87
regionPTR	.00 ± .00	13.5 ± 1.5	.60	.72 ± .25	**13.5** ± 1.0	.98
BRAL	**.45 ± .07**	**41.6**	**.76**	**.89 ± .07**	12.3	**.99**

from identifying it. BRAL succeeds ($F_1 = 0.45$) because point-wise persistence detects the branch tip regardless of cluster membership. On PBMC 3k, where the rare population forms a distinct island, regionPTR performs competitively ($F_1 = 0.72$), confirming the failure on Paul15 stems from the mismatch between region-based abstractions and branching manifold structures.

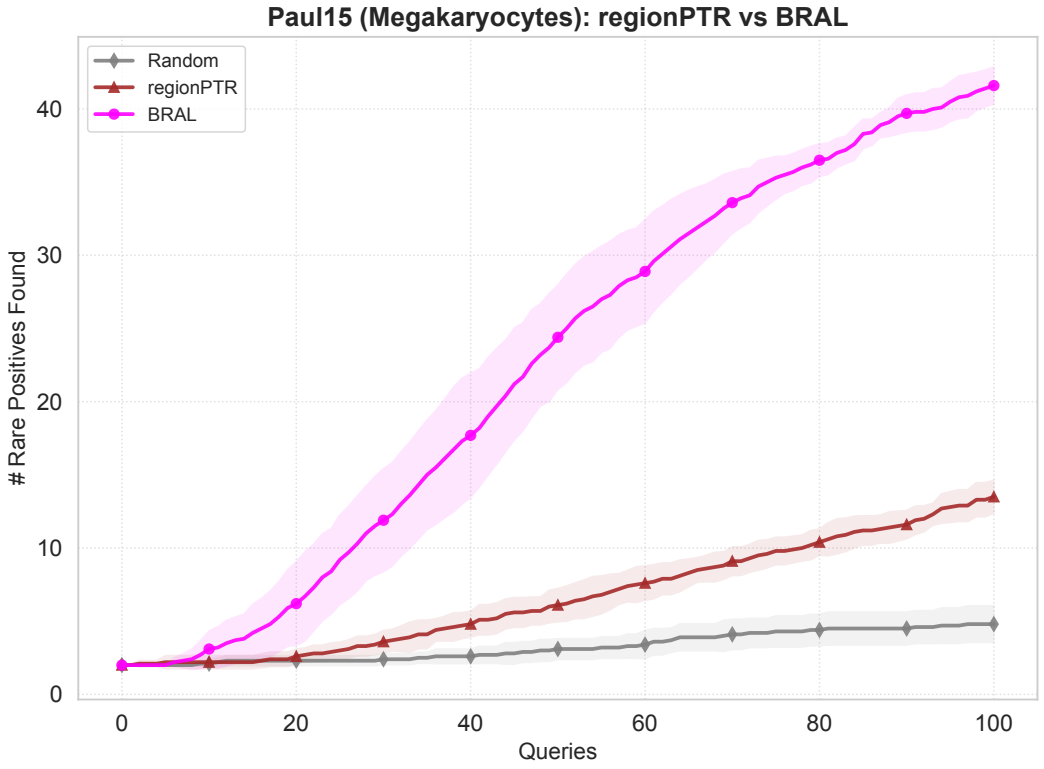

Figure 2: Paul15: Rare positive found comparison of regionPTR vs. BRAL over 100 queries.

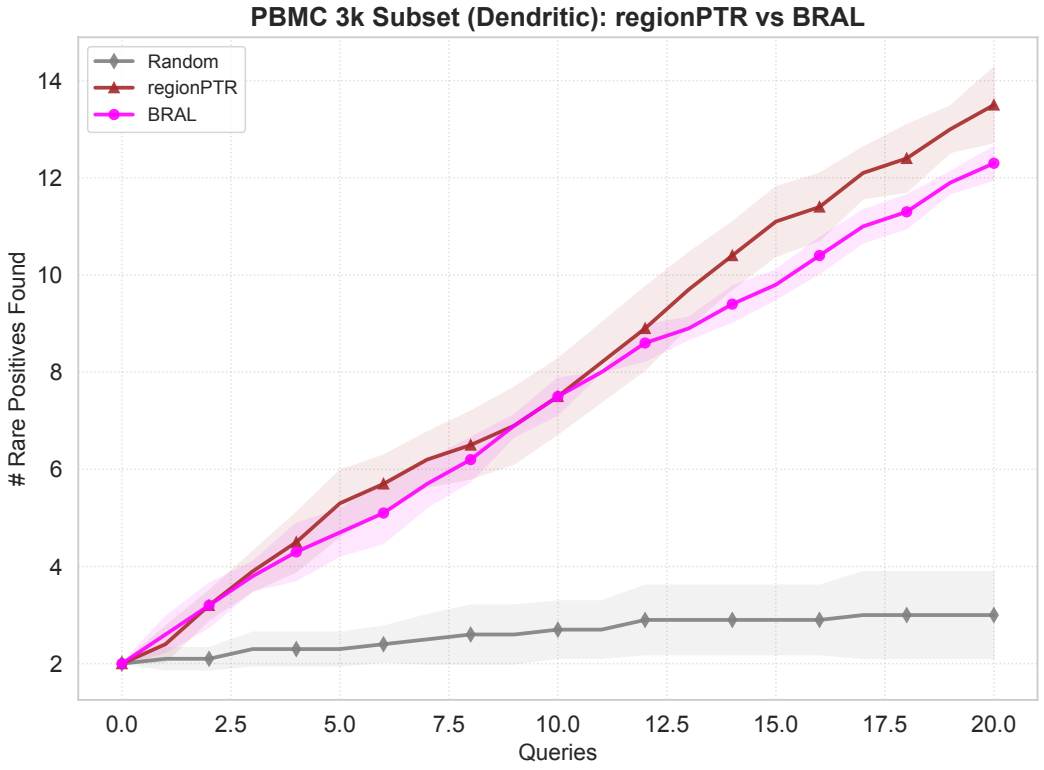

Figure 3: PBMC 3k subset: Rare positive found comparison of regionPTR vs. BRAL over 20 queries.

C Sensitivity to Hyperparameter λ

BRAL(Eq. 1) balances uncertainty and topological impact via the weighting parameter λ:

$$a_{\text{BRAL}}(x) = \underbrace{\widetilde{H}(x)}_{\text{exploitation}} + \lambda \cdot \underbrace{\widetilde{\Delta}_{\text{topo}}(x)}_{\text{topological signal}}, \tag{5}$$

We hypothesize that Dynamic Score Normalization (Eq. 2) renders the method robust to λ. We evaluated BRAL on the Synthetic Bifurcation dataset across $\lambda \in [0.1, 50]$.

Results: Performance is consistently high ($F_1 \approx 0.93$–0.94) for $\lambda \leq 1.0$, confirming that dynamic normalization brings the topological signal into a useful range. Beyond λ=2.0 (F_1=0.91), performance drops gradually with increasing variance (λ=5.0: F_1=0.89$\pm$0.09; λ=50.0: F_1=0.85$\pm$0.19), as the acquisition function increasingly ignores the classifier's uncertainty. We fixed $\lambda = 1.0$ for all main experiments.

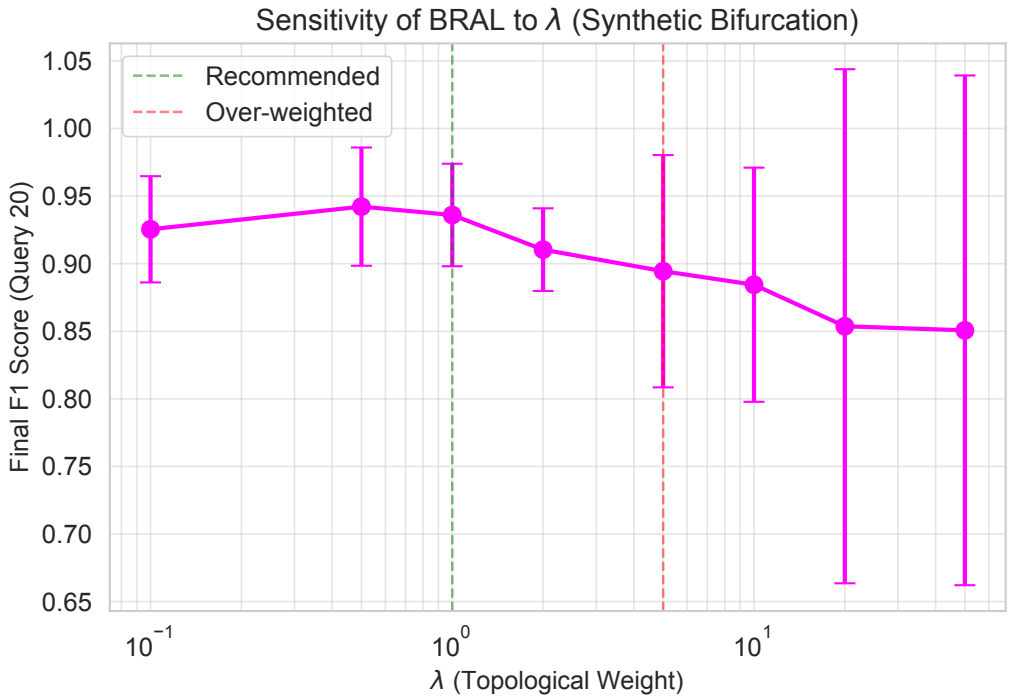

Figure 4: **Hyperparameter Sensitivity (λ).** Performance on the Synthetic Bifurcation task remains high and stable for $\lambda \leq 1$, confirming that Dynamic Normalization effectively scales the topological term. Performance degrades when topology is over-weighted ($\lambda \geq 5$).

D SHORTLIST MODE ABLATION

BRAL evaluates topological impact on a shortlist of 200 candidates to keep computation tractable. We compare three shortlisting strategies on the Synthetic Bifurcation dataset (N=20 queries, 10 seeds):

- **Hybrid** (proposed): 180 points with highest GP predictive entropy + 20 drawn uniformly at random.
- **TopUnc**: All 200 candidates selected by highest GP entropy.
- **PureRand**: All 200 candidates drawn uniformly at random.

Table 3: Shortlist mode ablation on synthetic bifurcation (mean ± std, 10 seeds). The hybrid shortlist achieves the best performance, confirming that both uncertainty-guided pre-filtering and random exploration contribute to BRAL's effectiveness.

Mode	F1	Discovery	AP
BRAL_Hybrid (proposed)	$\mathbf{0.94 \pm 0.03}$	$\mathbf{11.8 \pm 0.9}$	$\mathbf{0.97 \pm 0.01}$
BRAL_TopUnc	0.93 ± 0.04	11.6 ± 1.0	0.97 ± 0.02
BRAL_PureRand	0.82 ± 0.11	11.4 ± 1.9	0.94 ± 0.04
Random	0.67 ± 0.27	4.2 ± 1.3	0.85 ± 0.13
TopoOnly	0.85 ± 0.07	10.4 ± 0.9	0.94 ± 0.03

Hybrid and TopUnc achieve nearly identical F1 (0.94 vs. 0.93), with Hybrid showing lower variance (± 0.03 vs. ± 0.04). PureRand drops to $F_1 = 0.82$, confirming that uncertainty-guided pre-filtering

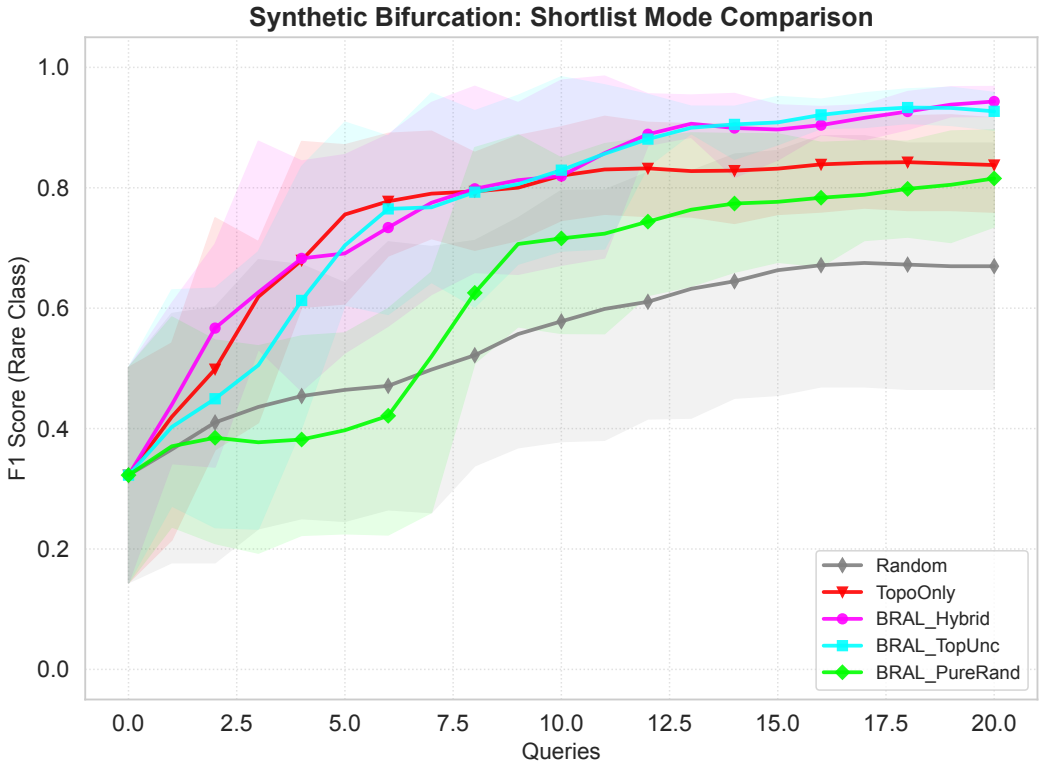

Figure 5: Shortlist mode comparison on Synthetic Bifurcation: F1 learning curves. Hybrid and TopUnc perform similarly, while PureRand is substantially weaker.

is essential. All BRAL variants outperform TopoOnly (0.84). We adopt Hybrid for all main experiments.

E Cold-Start Sensitivity

We evaluate BRAL and baselines on PBMC 3k subset under three cold-start conditions: **Standard** (2P + 3N), **Hard** (1P + 4N), and **Extreme** (0P + 5N, no positive examples).

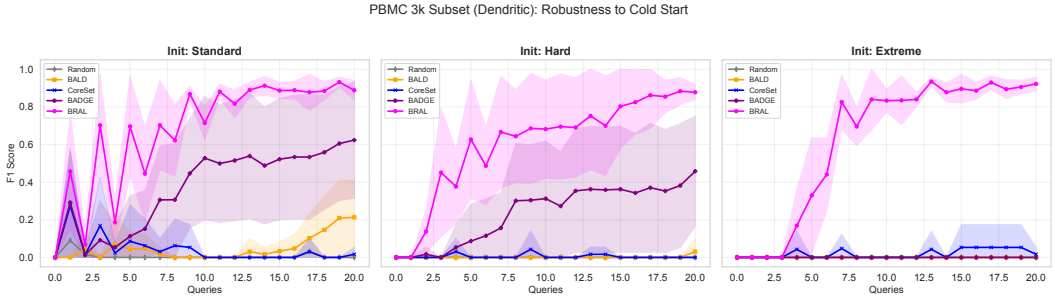

Figure 6: Cold-start sensitivity on PBMC 3k subset: F1 learning curves under Standard (2P+3N), Hard (1P+4N), and Extreme (0P+5N) initializations. BRAL maintains strong performance across all modes.

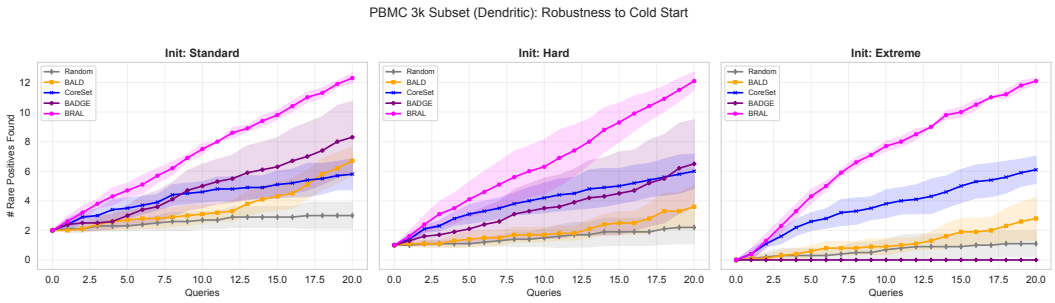

Figure 7: Cold-start sensitivity on PBMC 3k subset: cumulative rare cell discoveries under each initialization mode.

Table 4: Cold-start sensitivity on PBMC 3k subset (DC, 7.2%). Mean ± std over 10 seeds. BRAL maintains strong performance across all initialization modes, while baselines degrade severely.

Mode	Method	F1	Discovery
Standard (2P+3N)	Random	.00 ± .00	3.0 ± 1.2
	BALD	.21 ± .26	6.7 ± 1.3
	CoreSet	.02 ± .05	5.8 ± 1.4
	BADGE	.62 ± .41	8.3 ± 3.2
	BRAL	**.89 ± .07**	**12.3 ± 0.5**
Hard (1P+4N)	Random	.00 ± .00	2.2 ± 1.5
	BALD	.03 ± .09	3.6 ± 1.9
	CoreSet	.00 ± .00	6.0 ± 1.5
	BADGE	.46 ± .39	6.5 ± 4.0
	BRAL	**.88 ± .05**	**12.1 ± 0.8**
Extreme (0P+5N)	Random	.00 ± .00	1.1 ± 1.1
	BALD	.00 ± .00	2.8 ± 1.9
	CoreSet	.02 ± .05	6.1 ± 1.2
	BADGE	.00 ± .00	0.0 ± 0.0
	BRAL	**.92 ± .05**	**12.1 ± 0.3**

BRAL is robust across all cold-start conditions: $F_1 = 0.89$ (Standard), 0.88 (Hard), and 0.92 (Extreme). Under the Extreme setting, all baselines collapse to $F_1 = 0.00$; BADGE discovers zero rare cells because its gradient-based acquisition requires at least one positive example. BRAL's topological component operates independently of label information, detecting structural novelty in the data manifold regardless of whether the classifier has seen any positive examples.

F Computational Complexity

We analyzed the runtime cost of each acquisition function on the Paul15 dataset ($N \approx 2{,}730$, $D=50$, $N_q=100$). Table 5 reports the total wall-clock time for the full AL loop (100 queries), averaged across 10 seeds.

The overhead is dominated by the Rips persistence computation over the 200-point shortlist. The dominant cost shared by all methods is GP retraining ($\sim$0.5s per query). Table 6 reports per-query runtimes across all datasets.

Table 5: Total Runtime (seconds) for 100 queries on Paul15. BRAL is approximately 4–5× slower than standard baselines but the absolute per-query cost (∼2.7s) is negligible compared to biological oracle time.

Method	**Total Time (s)**	**Per Query (s)**
Random	54	~ 0.5
BALD	57	~ 0.6
CoreSet	60	~ 0.6
BADGE	62	~ 0.6
regionPTR	~ 55	~ 0.6
TopoOnly	~ 270	~ 2.7
BRAL (Ours)	274	~ 2.7

Table 6: Per-query runtime (seconds) across all datasets (10 seeds). The topological overhead is consistent across dataset sizes.

Method	**Synthetic** (N_q=20)	**Paul15** (N_q=100)	**PBMC 3k sub.** (N_q=20)	**PBMC 3k full** (N_q=100)
Random	0.47	0.54	0.37	~ 0.5
BALD	0.47	0.57	0.36	~ 0.5
CoreSet	0.47	0.60	0.48	~ 0.5
BADGE	0.47	0.62	0.36	~ 0.5
BRAL	3.95	2.74	3.16	~ 2.5

G BIO-DOMAIN BASELINES

We compare BRAL against domain-specific baselines that leverage biological prior knowledge or specialized model architectures:

- **RF_Entropy**: Random Forest (100 trees, depth 10, balanced class weights) with max-entropy acquisition (Geuenich et al., 2024).
- **MarkerExpr**: Selects cells with the highest expression of known lineage markers (PF4, PPBP, ITGA2B, GP5, TUBB1 for Mk; FCER1A, CST3, NPC2, HLA-DRA, HLA-DRB1 for DC).
- **SIMILAR**: FLQMI-based acquisition selecting points most similar to known positives via RBF kernel (Kothawade et al., 2021).

Table 7: Bio-domain Baseline Comparison (Mean ± Std, 10 seeds). On Paul15, BRAL is the best GP-based method. On PBMC 3k, BRAL achieves the highest F1 without requiring biological prior knowledge.

Method	**Disc@N**	**F1**	**AP**	**AUROC**
Paul15 (N=100)				
BRAL	$\mathbf{41.6 \pm 1.7}$	$\mathbf{0.45 \pm 0.07}$	$\mathbf{0.76 \pm 0.09}$	$\mathbf{0.99 \pm 0.01}$
MarkerExpr	30.5 ± 2.3	0.42 ± 0.10	0.54 ± 0.09	0.96 ± 0.01
RF_Entropy	38.1 ± 1.4	0.36 ± 0.09	0.72 ± 0.08	0.98 ± 0.01
SIMILAR	3.3 ± 1.3	0.00 ± 0.00	0.28 ± 0.14	0.93 ± 0.02
PBMC 3k subset (N=20)				
BRAL	12.3 ± 0.5	$\mathbf{0.89 \pm 0.07}$	$\mathbf{0.99 \pm 0.01}$	$\mathbf{1.00 \pm 0.00}$
RF_Entropy	10.8 ± 0.6	0.73 ± 0.24	0.99 ± 0.01	1.00 ± 0.00
MarkerExpr	$\mathbf{16.9 \pm 0.7}$	0.27 ± 0.12	0.98 ± 0.02	1.00 ± 0.00
SIMILAR	13.9 ± 5.8	0.16 ± 0.10	0.97 ± 0.03	1.00 ± 0.01

On Paul15, BRAL is the best GP-based method (F_1=0.45 which discovers the most rare cell (41.6) with highest AP and AUROC), outperforming RF_Entropy without domain knowledge. SIMILAR fails catastrophically (3.3 discoveries) due to exploitation of similarity to the 2 initial positives. On PBMC 3k, BRAL achieves the highest F1 (0.89). MarkerExpr discovers the most rare cells (16.9) but achieves low F1 (0.27), illustrating that high discovery alone is insufficient for boundary learning.

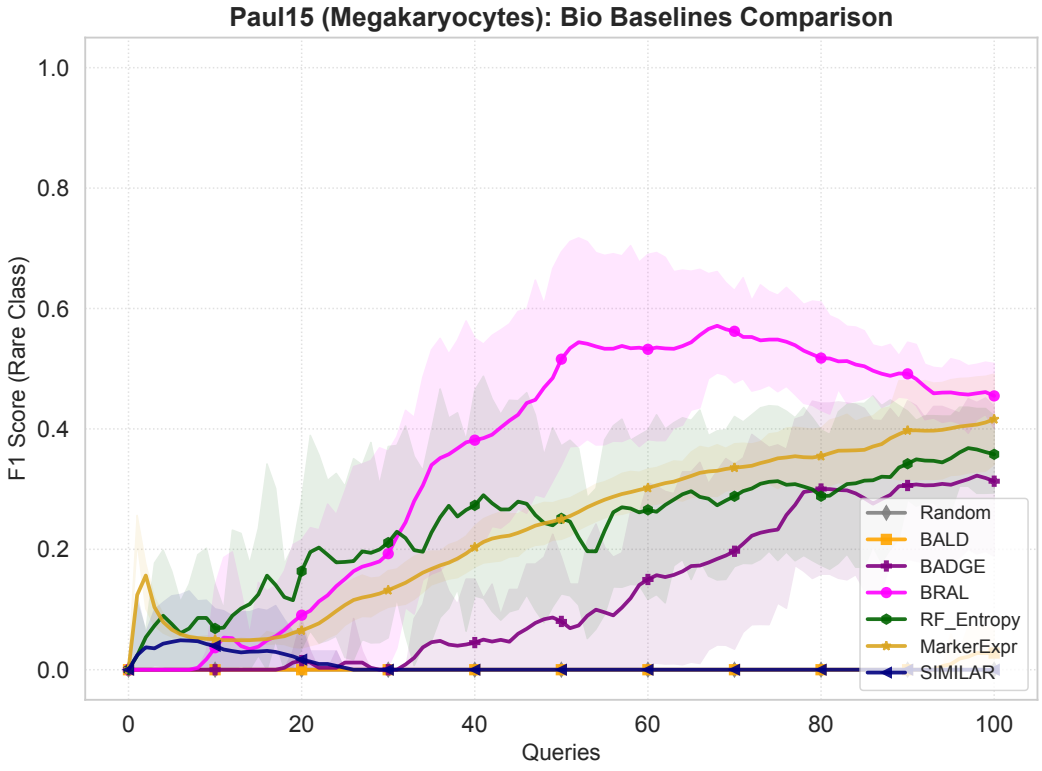

Figure 8: Bio-domain baselines on Paul15: F1 learning curves.

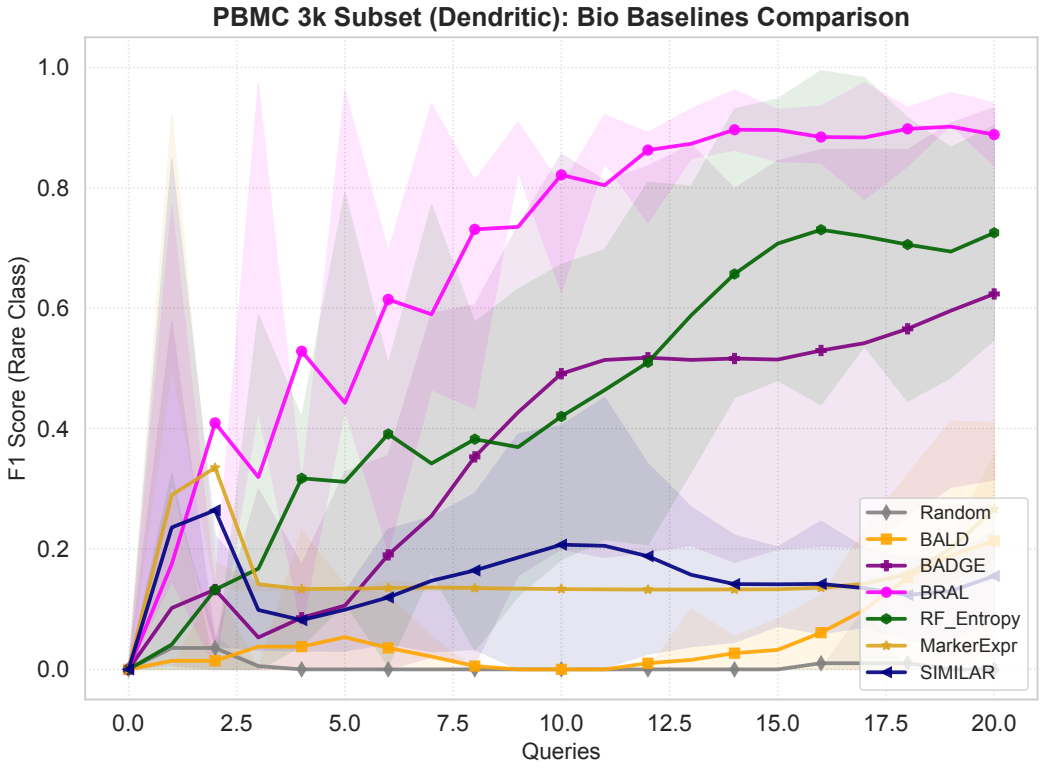

Figure 9: Bio-domain baselines on PBMC 3k subset: F1 learning curves.

H BUDGET SWEEP

We evaluated BRAL and baselines on Paul15 at three query budgets: $N \in \{20, 50, 100\}$, corresponding to approximately 1%, 2.5%, and 5% of the pool.

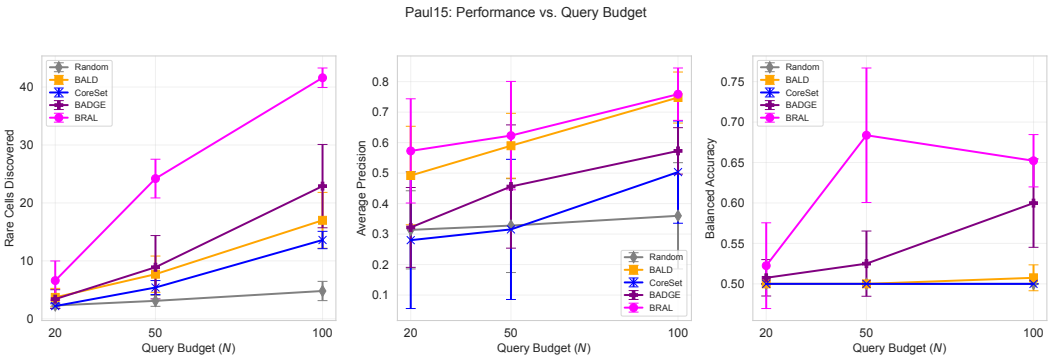

Figure 10: Budget sweep on Paul15. BRAL benefits most from increased budget, while baselines show limited improvement. Left: rare cells discovered. Center: average precision. Right: balanced accuracy.

At N=20, all methods struggle (BRAL: F_1=0.07, baselines: $F_1 \leq 0.03$). At N=50, BRAL separates (F_1=0.48, 24.2 discoveries) while baselines remain near zero. At N=100, the gap is substantial (F_1=0.45, 41.6 discoveries vs. BADGE's 0.31, 22.9). BRAL efficiently converts additional queries into rare-cell discoveries.

I MECHANISM VISUALIZATION: PERSISTENCE BARCODES

To illustrate how topological impact guides acquisition, we visualize the persistence barcodes before and after adding candidate points to the labeled set, using a small subset of the Synthetic Bifurcation dataset (n=100 points, 10 initial labeled).

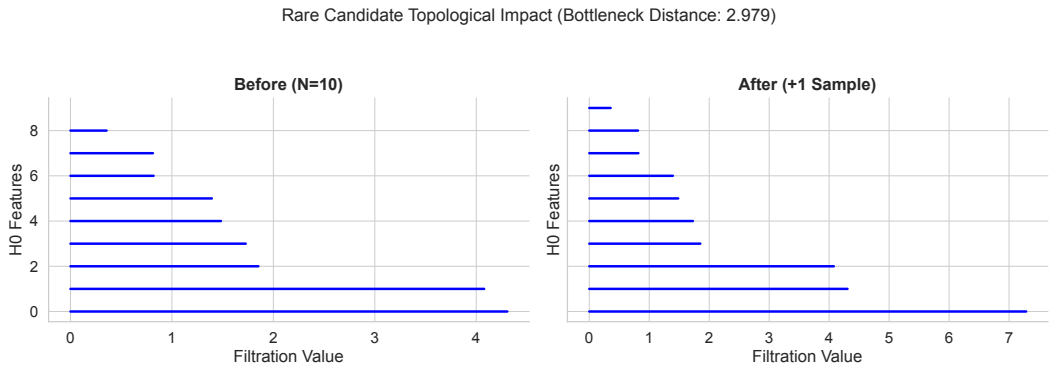

Figure 11: **Rare Candidate:** Adding a point from the rare branch significantly changes the H_0 persistence barcode, creating a new long-lived connected component. The bottleneck distance (topological impact) is large, triggering BRAL acquisition.

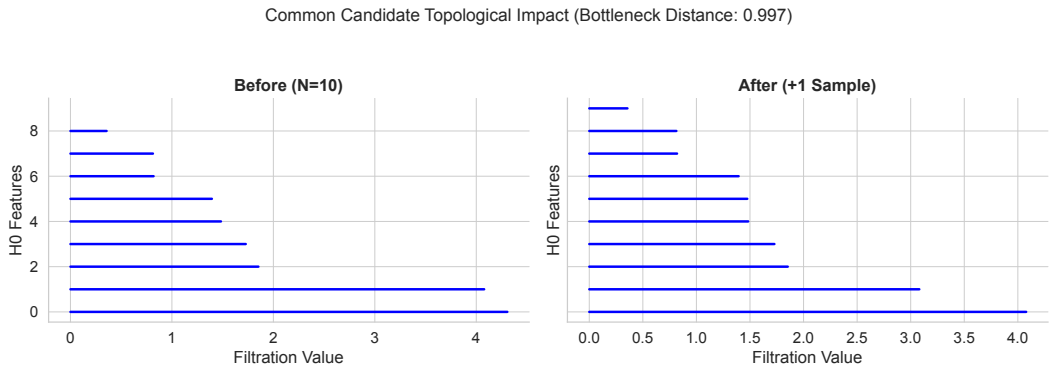

Figure 12: **Common Candidate:** Adding a point from the common class produces minimal change in the barcode. The bottleneck distance is small, so BRAL deprioritizes this candidate.

A rare-branch candidate (Figure 11) lies in an isolated region and its addition creates large changes in the H_0 persistence diagram. A common-class candidate (Figure 12) lies near existing labeled points, yielding a small bottleneck distance.

J THE COMPLEMENTARITY OF UNCERTAINTY AND TOPOLOGY

Table 8 summarizes the component ablation: neither uncertainty nor topology alone succeeds on real datasets.

TopoOnly achieves $F_1 = 0.00$ on both real datasets despite succeeding on the synthetic data (0.85). On real data with limited initial labels, the GP posterior is nearly flat, making the uncertainty-based shortlist effectively random; topology evaluated on an arbitrary subset of a high-dimensional manifold yields noisy scores. BALD discovers $\sim$17 rare cells on Paul15 but achieves $F_1 = 0.03$, indicating discovery without boundary learning—the discovered cells are spread too sparsely to shift the decision boundary past the 0.5 threshold.

Table 8: Component ablation: neither uncertainty alone (BALD) nor topology alone (TopoOnly) succeeds on real datasets. Only the combination (BRAL) achieves strong performance across all settings. Mean ± Std, 10 seeds.

Method	Synthetic	Paul15	PBMC 3k sub.
Random (baseline)	0.67 ± 0.27	0.00 ± 0.00	0.00 ± 0.00
BALD (uncertainty only)	0.78 ± 0.13	0.03 ± 0.06	0.21 ± 0.26
TopoOnly (topology only)	0.85 ± 0.07	0.00 ± 0.00	0.00 ± 0.00
BRAL (combined)	$\mathbf{0.94 \pm 0.03}$	$\mathbf{0.45 \pm 0.07}$	$\mathbf{0.89 \pm 0.07}$

K Full Multi-Metric Results

Table 9 presents all five evaluation metrics for the main experiments, complementing Table 1 in the main text.

Table 9: Full evaluation metrics across all datasets (Mean ± Std, 10 seeds). BRAL achieves the best or near-best performance on all metrics across Synthetic, Paul15, and PBMC 3k subset.

Method	Discovery@N	F1	AP	AUROC	Balanced Acc.
Synthetic Bifurcation (N=20, ∼9% rare)					
Random	4.2 ± 1.3	0.67 ± 0.27	0.85 ± 0.13	0.96 ± 0.03	0.79 ± 0.13
BALD	7.3 ± 1.5	0.78 ± 0.13	0.93 ± 0.02	0.98 ± 0.01	0.87 ± 0.08
CoreSet	7.3 ± 0.9	0.85 ± 0.06	0.93 ± 0.02	0.98 ± 0.01	0.87 ± 0.05
BADGE	10.9 ± 1.6	0.69 ± 0.22	0.88 ± 0.15	0.95 ± 0.06	0.85 ± 0.10
regionPTR	8.9 ± 0.9	0.74 ± 0.16	0.89 ± 0.06	0.96 ± 0.02	0.87 ± 0.05
TopoOnly	10.3 ± 0.8	0.85 ± 0.07	0.94 ± 0.02	0.98 ± 0.01	0.90 ± 0.03
BRAL	$\mathbf{11.8 \pm 0.9}$	$\mathbf{0.94 \pm 0.03}$	$\mathbf{0.97 \pm 0.01}$	$\mathbf{0.99 \pm 0.00}$	$\mathbf{0.95 \pm 0.03}$
Paul15 — Megakaryocytes (N=100, ∼2.5% rare)					
Random	4.8 ± 1.7	0.00 ± 0.00	0.36 ± 0.17	0.94 ± 0.02	0.50 ± 0.00
BALD	17.0 ± 4.8	0.03 ± 0.06	0.75 ± 0.08	0.98 ± 0.01	0.51 ± 0.02
CoreSet	13.6 ± 1.5	0.00 ± 0.00	0.50 ± 0.17	0.81 ± 0.14	0.50 ± 0.00
BADGE	22.9 ± 7.2	0.31 ± 0.16	0.57 ± 0.08	0.96 ± 0.01	0.60 ± 0.05
regionPTR	13.5 ± 1.5	0.00 ± 0.00	0.60 ± 0.18	0.90 ± 0.12	0.50 ± 0.00
TopoOnly	6.5 ± 1.2	0.00 ± 0.00	0.64 ± 0.09	0.97 ± 0.01	0.50 ± 0.00
BRAL	$\mathbf{41.6 \pm 1.7}$	$\mathbf{0.45 \pm 0.07}$	$\mathbf{0.76 \pm 0.09}$	$\mathbf{0.99 \pm 0.01}$	$\mathbf{0.65 \pm 0.03}$
PBMC 3k Subset — Dendritic (N=20, ∼7.2% rare)					
Random	3.0 ± 1.2	0.00 ± 0.00	0.87 ± 0.24	0.96 ± 0.10	0.50 ± 0.00
BALD	6.7 ± 1.3	0.21 ± 0.26	0.99 ± 0.02	1.00 ± 0.00	0.57 ± 0.09
CoreSet	5.8 ± 1.4	0.02 ± 0.05	0.94 ± 0.08	0.99 ± 0.02	0.50 ± 0.01
BADGE	8.3 ± 3.2	0.62 ± 0.41	0.95 ± 0.08	0.99 ± 0.01	0.78 ± 0.19
regionPTR	13.5 ± 1.0	0.72 ± 0.25	0.98 ± 0.02	1.00 ± 0.00	0.93 ± 0.06
TopoOnly	2.9 ± 0.8	0.00 ± 0.00	0.85 ± 0.22	0.96 ± 0.07	0.50 ± 0.00
BRAL	$\mathbf{12.3 \pm 0.5}$	$\mathbf{0.89 \pm 0.07}$	$\mathbf{0.99 \pm 0.01}$	$\mathbf{1.00 \pm 0.00}$	0.94 ± 0.06

L PBMC 3k Full Dataset Analysis

The full PBMC 3k dataset (∼2,700 cells, all cell types, Dendritic cells ∼1.2% of population) represents the most challenging setting. Unlike the PBMC 3k subset (2 cell types), the full dataset collapses 8+ cell types into the "common" class. Table 10 reports results at N=100 queries.

This is the only dataset where BRAL does not achieve the highest final F1. BADGE (0.69) outperforms BRAL (0.64) despite BRAL discovering more rare cells (26.0 vs. 24.4). The gap is attributable to majority-class heterogeneity: the "common" class contains 7+ cell types, and BRAL's

Table 10: PBMC 3k Full results (N=100, Mean ± Std, 10 seeds). BADGE achieves the highest final F1, though BRAL discovers more rare cells and achieves higher AP and AUROC.

Method	Disc.	F1	AP	AUROC	Bal Acc
Random	3.0	0.00	0.97	1.00	0.50
BALD	12.5	0.32	0.97	1.00	0.61
CoreSet	3.7	0.00	0.79	0.96	0.50
BADGE	24.4	**0.69**	0.99	1.00	**0.77**
regionPTR	5.6	0.00	0.92	0.99	0.50
TopoOnly	2.9	0.00	0.92	1.00	0.50
BRAL	**26.0**	0.64	**0.99**	**1.00**	0.74

aggressive rare-cell acquisition creates a labeled set that is ∼48% rare by query 20, leaving insufficient majority-class diversity for the GP to distinguish phenotypically similar sub-types from Dendritic cells. BRAL reaches a peak F1 of ∼0.93 at query 24, then declines as continued acquisition after the rare class is fully discovered dilutes the labeled set. Despite the lower final F1, BRAL achieves perfect discovery ($26/26$ on all seeds) and the highest AP and AUROC, indicating the F1 gap is driven by threshold sensitivity rather than ranking quality.

M Baseline Descriptions

For completeness, we provide detailed descriptions of all baseline acquisition functions.

Random. Selects a point uniformly at random from the remaining pool $\mathcal{U}_t$.

BALD. Bayesian Active Learning by Disagreement (Houlsby et al., 2011). We implement BALD with the GP classifier using Monte Carlo sampling (S=25 posterior samples):

$$a_{\text{BALD}}(x) \approx H[\bar{p}(x)] - \tfrac{1}{S}\sum_{s=1}^{S} H[p_s(x)],$$

where $p_s(x)$ is the predictive probability from the s-th posterior sample and $\bar{p}(x)$ is the mean prediction.

CoreSet. Greedy k-center approach (Sener & Savarese, 2018) that selects the pool point maximizing distance to the nearest labeled point:

$$a_{\text{CoreSet}}(x) = \min_{z \in X_{\text{train}}^{(t)}} \|x - z\|_2.$$

BADGE. Batch Active Learning by Diverse Gradient Embeddings (Ash et al., 2020). We adapt BADGE to the GP setting by computing gradient norms of the variational ELBO loss with respect to input features.

regionPTR. A region-based topological baseline inspired by Proper Topological Regions (Hadjadj et al., 2023). Detailed implementation in Appendix B.

TopoOnly. An ablation that uses only topological impact ($a_{\text{TopoOnly}}(x) = \Delta_{\text{topo}}(x; S)$), evaluated on the same uncertainty-based shortlist as BRAL.

N Vietoris–Rips Persistent H_0

Given a finite point cloud $S \subset \mathbb{R}^d$ and a scale parameter $r \geq 0$, the Vietoris–Rips complex $\text{VR}_r(S)$ is the abstract simplicial complex whose k-simplices correspond to subsets of S of diameter at most $2r$. Equivalently, a simplex is included whenever all pairwise distances between its vertices are $\leq 2r$.

As r increases from 0 to ∞, the complexes $\mathrm{VR}_r(S)$ form a filtration

$$\mathrm{VR}_0(S) \subseteq \mathrm{VR}_{r_1}(S) \subseteq \mathrm{VR}_{r_2}(S) \subseteq \cdots,$$

encoding the multi-scale connectivity of S. The zeroth homology group $H_0(\mathrm{VR}_r(S))$ records the number of connected components at scale r. Initially, each point is its own component; as r grows, components merge but never split.

Persistent H_0 tracks these merge events across scales via a persistence diagram $D_0(S)$. Each point in $D_0(S)$ has coordinates (b, d), where b is the scale at which a component is born and d is the scale at which it merges with another. Long-lived points ($d - b$ large) represent robust separation in the data, while short-lived points typically reflect noise.

To compare the topology before and after adding a candidate x, we use the bottleneck distance

$$d_B\big(D_0(S), D_0(S \cup \{x\})\big),$$

which measures the largest change in birth or death times under an optimal matching of points between diagrams. Stability results guarantee that small perturbations of S induce small changes in $D_0(S)$, making d_B a reliable notion of topological impact. In BRAL, a large bottleneck distance indicates that x creates or significantly shifts a connected component, suggesting that it lies in a topologically distinct region of the data manifold.