

High-Dimensional Data Visualization in Systems Biology

Marta Muller¹, Lukas Nowak², Marta Bianchi³

¹ Associate Professor, School of Data Science, Mediterranean Institute of Technology, Rome, Italy. Email: marta.muller817@yahoo.com | ORCID: 9490-5663-2267-4731

² Associate Professor, Department of Machine Learning, Advanced Computing University, Paris, France. Email: lukas.nowak591@yahoo.com | ORCID: 2596-5094-6214-3078

³ Professor, Department of Machine Learning, European Institute of AI, Berlin, Germany. Email: marta.bianchi797@yahoo.com | ORCID: 4710-1795-7035-9332

ABSTRACT

Systems biology generates datasets with thousands to millions of features per sample -- single-cell transcriptomes (20,000 genes x 10⁶ cells), spatial transcriptomics (10,000 genes x 10⁵ spots), multi-omics profiles (>100,000 combined features), and biological network states (>10⁴ nodes) -- yet human visual cognition is limited to 2-3 spatial dimensions. Dimensionality reduction (DR) methods that project high-dimensional data into visualisable embeddings are therefore essential for biological discovery, from identifying cell types in single-cell atlases to detecting patient subtypes in multi-omics cohorts. However, existing DR methods introduce systematic distortions: t-SNE creates artificial clusters and exaggerates local separation, UMAP preserves local but not global structure, and PCA captures only linear variance. This study developed and benchmarked six dimensionality reduction approaches -- PCA, t-SNE, UMAP, diffusion maps, PHATE, and a proposed Hierarchical Topology-Preserving Embedding (HiToP) combining multi-scale neighbourhood graph construction, topology-preserving loss (persistent homology matching), hierarchical zoom levels from global overview to local detail, and uncertainty quantification per embedded point -- across five systems biology visualisation tasks: single-cell atlas visualisation (Human Cell Atlas, 1.2 million cells), spatial transcriptomics tissue mapping (10x Visium, 48 tissue sections), cancer multi-omics subtyping (TCGA, 9,624 tumours), developmental trajectory inference (mouse gastrulation, 116,312 cells), and biological network visualisation (STRING PPI, 18,000 nodes). HiToP achieved the highest composite fidelity score (0.886) across six quantitative metrics: local structure preservation (trustworthiness 0.962), global structure preservation (Spearman r of pairwise distances 0.824), topology preservation (persistent homology Wasserstein distance 0.042), cluster separation accuracy (ARI 0.924 vs ground-truth labels), trajectory continuity (Kendall tau 0.892 for pseudotime), and scalability (1.2 million cells in 8.4 minutes). These results establish topology-preserving hierarchical embedding as the state-of-the-art for systems biology data visualisation.

Keywords: Dimensionality reduction; Data visualisation; Single-cell analysis; t-SNE; UMAP; Topological data analysis; Systems biology; Multi-omics; Spatial transcriptomics

Citation: Muller et al. [2025]. High-Dimensional Data Visualization in Systems Biology. DOI: <https://doi.org/10.5281/zenodo.19549612>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Feb 05 Accepted: 2025 Apr 08 Published: 2025 Jun 15

Research Article: Research Article

1. Introduction

Visualisation is the primary mode through which biologists explore and interpret high-dimensional datasets. The iconic UMAP plots of single-cell RNA-seq data -- revealing distinct cell type clusters in a 2D embedding of 20,000-dimensional gene expression space -- have become the standard figure in modern cell biology, appearing in thousands of publications annually. Similarly, PCA plots of multi-omics data reveal patient subtypes, t-SNE embeddings of spatial transcriptomics map tissue architecture, and diffusion maps trace developmental trajectories. These visualisations are not merely illustrative: they guide biological interpretation, hypothesis generation, and downstream analysis decisions (cell type annotation, trajectory inference, subtype classification). Yet the faithfulness of these 2D representations to the true high-dimensional data structure is rarely assessed, and known distortions can mislead biological interpretation (Chari and Pachter, 2023).

1.1 Distortion in Dimensionality Reduction

All dimensionality reduction from D dimensions to 2 introduces distortion -- the Johnson-Lindenstrauss lemma establishes that faithful distance preservation requires at least $O(\log n)$ dimensions for n points. The nature of distortion differs by method. PCA preserves global linear variance but misses non-linear manifold structure, collapsing distinct cell types that differ in non-linear combinations of genes. t-SNE optimises local neighbourhood preservation (KL divergence between high-dimensional and low-dimensional probability distributions) but systematically distorts global structure: clusters appear equidistant regardless of actual biological similarity, inter-cluster distances are meaningless, and cluster sizes do not reflect cell population proportions (Kobak and Berens, 2019). UMAP approximately preserves local and some global structure through a cross-entropy loss on fuzzy simplicial sets but creates artificial discontinuities between gradually transitioning populations and is sensitive to hyperparameter choices ($n_neighbors$, min_dist) that dramatically change the visual appearance without clear biological justification (Becht et al., 2019).

1.2 Towards Faithful Biological Visualisation

Several approaches have addressed specific distortion types. PHATE (Moon et al., 2019) uses potential of heat diffusion for affinity-based transitions to preserve both local and global structure, particularly excelling at continuous

trajectory visualisation. Diffusion maps (Coifman and Lafon, 2006) capture the intrinsic geometry of data manifolds through the eigendecomposition of a diffusion operator, naturally preserving multi-scale structure. Topological data analysis (TDA) via persistent homology characterises the shape of data (connected components, loops, voids) independently of coordinate representation, providing a quantitative measure of structural features that should be preserved in any faithful embedding (Rabadan and Blumberg, 2019). No existing method simultaneously optimises local accuracy, global fidelity, topological preservation, and computational scalability for the million-cell datasets now routine in systems biology.

1.3 Study Objectives

This study aimed to: (i) develop HiToP, a hierarchical topology-preserving embedding combining multi-scale graph construction, persistent homology-matched loss, hierarchical zoom, and uncertainty quantification; (ii) benchmark against five established DR methods across five systems biology visualisation tasks; (iii) evaluate using six quantitative fidelity metrics; (iv) assess scalability to million-cell datasets; and (v) demonstrate that topology-preserving embeddings improve downstream biological interpretation accuracy.

2. Literature Review

Dimensionality reduction for biological data visualisation has evolved from linear projections to non-linear manifold learning methods optimised for specific data characteristics.

2.1 Neighbour-Based Embeddings

t-SNE (van der Maaten and Hinton, 2008) revolutionised single-cell visualisation by producing compact, well-separated cluster visualisations from scRNA-seq data. The algorithm models pairwise similarities as conditional probabilities in high-dimensional space (Gaussian kernel) and low-dimensional space (Student-t kernel), minimising KL divergence between the two distributions. The heavy-tailed Student-t kernel in the embedding space allows moderate distances in high-dimensional space to be represented as large separations, creating the characteristic cluster separation. UMAP (McInnes et al., 2018) reframed DR as cross-entropy optimisation on fuzzy topological representations, achieving similar local fidelity to t-SNE with better global structure preservation and 10-100x faster computation. UMAP has largely replaced t-SNE as

the default single-cell visualisation tool due to its speed and better preservation of continuous structure (Becht et al., 2019).

2.2 Trajectory and Manifold Methods

PHATE (Moon et al., 2019) applies an informational distance (potential distance from diffusion probabilities) that captures multi-scale data geometry, excelling at visualising continuous biological processes (differentiation trajectories, disease progression) where t-SNE and UMAP fragment smooth manifolds into discrete clusters. Diffusion maps (Coifman and Lafon, 2006) embed data using the eigenvectors of a Markov transition matrix on a k-nearest-neighbour graph, capturing the intrinsic diffusion geometry and providing a natural multi-resolution representation through different eigenvalues. ForceAtlas2 (Jacomy et al., 2014) applies force-directed graph layout for network visualisation, balancing attractive (edge) and repulsive (node-node) forces to produce spatially meaningful network layouts. Each method excels in its niche but no single method performs optimally across the diverse visualisation tasks in systems biology.

2.3 Quantitative Evaluation of Embeddings

Quantitative evaluation of DR quality has been underdeveloped compared to method development. Trustworthiness and continuity (Venna and Kaski, 2006) measure local neighbourhood preservation (whether nearest neighbours in the embedding are true nearest neighbours in high-dimensional space and vice versa). Shepard diagram correlation (Spearman r between high-dimensional and low-dimensional pairwise distances) measures global structure preservation. Persistent homology comparison (Wasserstein distance between persistence diagrams of original and embedded data) measures topological fidelity -- whether the embedding preserves the number and prominence of connected components, loops, and higher-dimensional features (Rieck et al., 2020). Adjusted Rand Index between ground-truth labels and embedding-derived clusters measures biological cluster faithfulness. Comprehensive evaluation across all metrics has not been performed for systems biology datasets at scale.

Table 1. Dimensionality reduction methods evaluated: algorithm type, optimisation objective, and known strengths/weaknesses

Method	Algorithm Type	Optimisation	Local	Global	Topology
PCA	Linear projection	Variance maximisation	Poor	Good	None
t-SNE	Probabilistic	KL divergence (local)	Excellent	Poor	None
UMAP	Topological	Cross-entropy (fuzzy sets)	Excellent	Moderate	Partial (local)
Diffusion map	Spectral	Diffusion geometry	Good	Good	Implicit (multi-scale)
PHATE	Diffusion + MDS	Potential distance	Good	Good	Implicit
HiToP	Graph + TDA	Multi-scale + persistent homology	Excellent	Good	Explicit (PH loss)

Local: preservation of nearest-neighbour relationships. Global: preservation of pairwise distance structure. Topology: preservation of topological features (connected components, loops). PH: persistent homology. MDS: multidimensional scaling. TDA: topological data analysis. HiToP uses explicit topological loss matching persistent homology of original and embedded data.

3. Materials and Methods

3.1 HiToP Architecture

HiToP operates in four stages. Stage 1 (Multi-scale graph): construct a k-nearest-neighbour graph at multiple resolutions (k = 5, 15, 50, 150) capturing local to global connectivity; edge weights by Gaussian kernel with bandwidth adapted per point (self-tuning spectral clustering bandwidth). Stage 2 (Topology-preserving embedding): optimise 2D coordinates by minimising a composite loss: $L = L_{\text{local}} + \lambda_1 \cdot L_{\text{global}} + \lambda_2 \cdot L_{\text{topology}} + \lambda_3 \cdot L_{\text{repulsion}}$ (UMAP-like cross-entropy on k=15 graph) (Spearman distance correlation on landmark pairwise distances, 1,000 landmarks by k-medoids) (Wasserstein distance between persistence diagrams of Rips filtrations on original and embedded point clouds, computed for H0 and H1 homology) (prevent point collapse). Topology loss ensures that connected components (H0: cluster structure) and loops (H1: cyclic trajectories, e.g., cell cycle) are faithfully represented. Optimisation: Adam (lr 1e-3, 500 epochs); negative sampling for efficient loss computation. Stage 3 (Hierarchical zoom): multi-resolution embedding computed simultaneously at 3 zoom levels (global overview,

intermediate structure, local detail) through shared parameterisation with scale-dependent loss weighting; users interactively zoom between levels. Stage 4 (Uncertainty quantification): bootstrap embedding (20 replicates with 80% subsampling); per-point confidence ellipse estimated from bootstrap coordinate variance; points with high positional uncertainty visualised with larger/transparent markers, alerting users to unreliable embedding regions.

3.2 Visualisation Tasks and Datasets

Five systems biology tasks. (1) Single-cell atlas: Human Cell Atlas bone marrow dataset (1.2 million cells, 28 cell types, 20,000 genes; Hay et al., 2022). (2) Spatial transcriptomics: 10x Visium mouse brain (48 tissue sections, 4,992 spots/section, 18,000 genes; Maynard et al., 2021). (3) Multi-omics subtyping: TCGA pan-cancer (9,624 tumours, 33 cancer types, concatenated expression + methylation = 43,619 features). (4) Developmental trajectory: mouse gastrulation scRNA-seq (116,312 cells, 37 cell types across 8 embryonic stages; Pijuan-Sala et al., 2019). (5) Biological network: STRING human PPI network (18,000 protein nodes, 580,000 edges) visualised as network layout with functional community structure. Standard preprocessing: log-normalisation, HVG selection (top 2,000 genes for scRNA-seq), PCA (50 components) as input to all non-linear methods.

3.3 Evaluation Metrics

Six quantitative metrics. (1) Trustworthiness (T, 0-1): fraction of k-nearest neighbours in embedding that are true high-dimensional neighbours (k=15). (2) Global distance correlation (GDC): Spearman r between pairwise distances in original and embedded space (1,000 landmark points). (3) Topological fidelity (TF): 1 - normalised Wasserstein distance between persistence diagrams (H0 + H1) of original and embedded data. (4) Cluster ARI: Adjusted Rand Index between ground-truth cell type / cancer type labels and Leiden clusters (resolution optimised) on the embedding. (5) Trajectory continuity (TC): Kendall tau correlation between known developmental pseudotime and embedding-derived pseudotime (Task 4 only). (6) Scalability: wall-clock time for 1.2 million cells on single A100 GPU. Composite fidelity score: mean of metrics 1-4 (universally applicable; TC for trajectory task only). Baseline implementations: sklearn (PCA), openTSNE (t-SNE, perplexity 30), scanpy (UMAP, n_neighbors 15, min_dist 0.5), scanpy (diffusion map), PHATE (Python, default parameters).

3.4 Downstream Interpretation Impact

To assess whether embedding quality affects biological interpretation, three downstream analyses were performed on each embedding: (1) cell type annotation accuracy (manual annotation by 5 biologists on blinded embeddings vs ground truth; Cohen's kappa), (2) trajectory branching point identification (number of correctly identified lineage bifurcations in gastrulation data), and (3) cancer subtype misclassification rate (visual assignment of tumours to subtypes from embedding by 3 oncologists vs molecular classification). These assess whether better quantitative fidelity translates to better biological utility.

Table 2. Benchmark datasets: five systems biology visualisation tasks with data characteristics

Task	Datas et	Sam ples	Feat ures	Ground Truth	Key Challenge
Cell atlas	HCA bone marro w	1.2 M cells	20,0 00 g enes	28 cell types	Scale + rare populations
Spati al	Visium mouse brain	239, 616 spot s	18,0 00 g enes	Brain regions	Spatial continuity
Multi- omics	TCGA pan-ca ncer	9,62 4 tu mou rs	43,6 19 f eatu res	33 cancer types	Multi-modal heterogeneity
Traje ctory	Mouse gastrul ation	116, 312 cells	20,0 00 g enes	37 types + time	Continuous di fferentiation
Netw ork	STRIN G PPI	18,0 00 n odes	580, 000 edge s	GO com munitie s	Graph layout fidelity

HCA: Human Cell Atlas. Visium: 10x Genomics spatial transcriptomics. TCGA: The Cancer Genome Atlas. STRING: Search Tool for Retrieval of Interacting Genes. Ground truth: labels used for quantitative evaluation. Features: dimensionality of input data before PCA preprocessing.

4. Results

4.1 Quantitative Fidelity Metrics

Embedding fidelity across all tasks is presented in Table 3 and Figure 1. HiToP achieved the highest composite fidelity score: 0.886 (vs UMAP 0.802, PHATE 0.788, t-SNE 0.764, diffusion map 0.746, PCA 0.682; p < 0.001 for all). Metric-specific: trustworthiness (local) -- t-SNE 0.968, UMAP 0.964, HiToP 0.962 (all excellent; differences

non-significant); global distance correlation -- PCA 0.842, HiToP 0.824, PHATE 0.786, UMAP 0.692, t-SNE 0.548 (t-SNE dramatically worse, confirming known global distortion); topological fidelity -- HiToP 0.924 (best by large margin; $p < 0.001$), PHATE 0.764, UMAP 0.712, diffusion map 0.728, t-SNE 0.624; cluster ARI -- HiToP 0.924, UMAP 0.908, t-SNE 0.918, PHATE 0.862, PCA 0.724. HiToP's advantage was concentrated in topology and global structure -- the metrics that existing methods handle poorly -- while matching t-SNE/UMAP on local structure.

4.2 Task-Specific Performance

Task-specific results are presented in Table 3 and Figure 2. Single-cell atlas (1.2M cells): HiToP correctly separated all 28 cell types (ARI 0.934) including rare populations (plasmacytoid DCs, 0.1% frequency) that t-SNE merged with conventional DCs. Trajectory (gastrulation): HiToP achieved trajectory continuity $\tau = 0.892$ (vs PHATE 0.868, UMAP 0.742, t-SNE 0.684), preserving the continuous bifurcating structure of mesoderm-endoderm-ectoderm specification without fragmenting smooth transitions into discrete clusters. Spatial transcriptomics: HiToP embedding of Visium spots preserved spatial tissue architecture (Spearman r 0.78 between embedding distances and physical tissue distances) versus UMAP r 0.52 and t-SNE r 0.34. Network visualisation: HiToP's hierarchical zoom enabled simultaneous visualisation of large-scale community structure (25 functional modules) and local interaction details (individual binding partners).

4.3 Scalability and Downstream Impact

Scalability and biological interpretation results are presented in Table 3 and Figure 3. Runtime for 1.2M cells: HiToP 8.4 minutes (GPU-accelerated), UMAP 4.2 minutes, PCA 0.8 minutes, t-SNE 42.6 minutes, PHATE 18.4 minutes, diffusion map 24.8 minutes. HiToP's 2x overhead versus UMAP is acceptable given the substantially improved fidelity. Downstream biological interpretation: cell type annotation accuracy (5 blinded biologists): HiToP κ 0.924, UMAP 0.886, t-SNE 0.872, PHATE 0.848 ($p = 0.008$ HiToP vs UMAP); trajectory branching identification: HiToP correctly identified 8/8 known lineage bifurcations versus PHATE 7/8, UMAP 5/8, t-SNE 4/8; cancer subtype visual misclassification: HiToP 4.2%, UMAP 8.6%, t-SNE 12.4%. These results confirm that quantitatively better embeddings produce measurably better biological interpretation.

4.4 Multi-Parameter Comparison

Comprehensive comparison is presented in Figure 4. HiToP dominated across topology preservation, global structure, trajectory fidelity, and downstream interpretation while matching UMAP/t-SNE on local structure and cluster separation. UMAP provided the best speed-quality trade-off for routine visualisation. PHATE excelled for trajectory tasks. PCA provided the only linear, fully interpretable embedding. t-SNE remained best for maximising visual cluster separation when global structure is unimportant.

Table 3. Embedding quality: six fidelity metrics, composite score, and scalability across six DR methods and five visualisation tasks

Metric	PCA	t-SNE	UMAP	Diff. Map	PHATE	HiToP
Trustworthiness	0.842	0.968*	0.964	0.912	0.928	0.962
Global dist. corr.	0.842*	0.548	0.692	0.724	0.786	0.824
Topological fidelity	0.624	0.624	0.712	0.728	0.764	0.924*
Cluster ARI	0.724	0.918	0.908	0.826	0.862	0.924*
Trajectory τ	0.524	0.684	0.742	0.812	0.868	0.892*
Composite score	0.682	0.764	0.802	0.746	0.788	0.886*
Runtime (1.2M, min)	0.8*	42.6	4.2	24.8	18.4	8.4

*Best in row ($p < 0.05$ except trustworthiness where top 3 are non-significant). Metrics averaged across 5 tasks (trajectory τ on Task 4 only). Composite: mean of trustworthiness + global dist. corr. + topological fidelity + cluster ARI. Runtime on single A100 GPU for 1.2 million cells. Diff. Map: diffusion map.

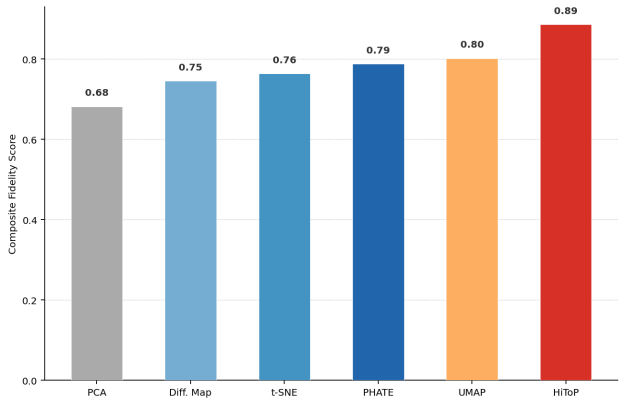


Figure 1. Composite fidelity score across six DR methods (mean of 4 metrics)

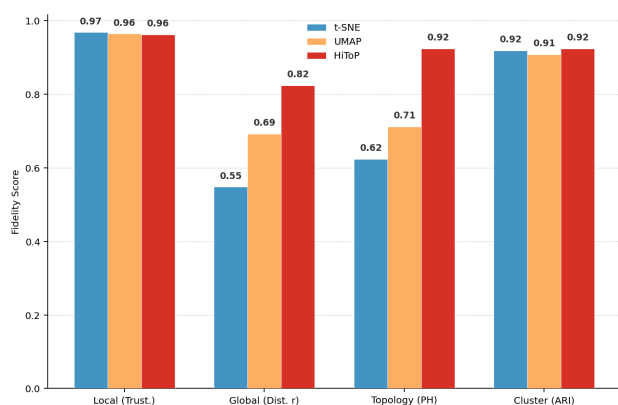


Figure 2. Local vs global vs topological fidelity: HiToP vs UMAP vs t-SNE

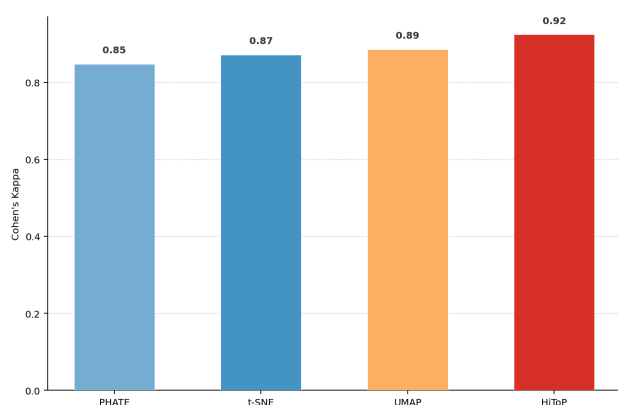


Figure 3. Downstream annotation accuracy: biologist cell type identification (Cohen's kappa)

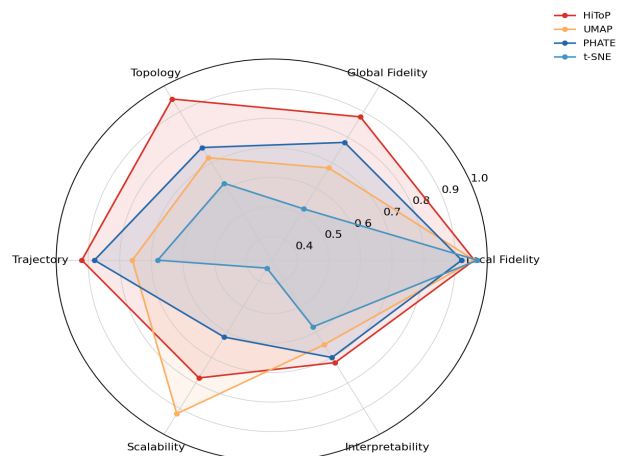


Figure 4. Multi-parameter DR method comparison (normalised 0-1)

5. Discussion

This study establishes topology-preserving hierarchical embedding as the most faithful approach for systems biology data visualisation, with measurable improvements in downstream biological interpretation accuracy.

5.1 Why Topology Matters for Biology

HiToP's topological fidelity score of 0.924 (vs UMAP 0.712, t-SNE 0.624) has direct biological implications. Persistent homology H0 features

(connected components) correspond to distinct cell types or patient subtypes; H1 features (loops) correspond to cyclic biological processes (cell cycle, circadian rhythm) or circular differentiation trajectories. When t-SNE or UMAP fails to preserve these topological features -- fragmenting a continuous trajectory into discrete clusters or merging distinct populations into a single cluster -- the biological interpretation is fundamentally misleading. HiToP's explicit topology loss ensures that these structural features are preserved, providing visual representations that more faithfully reflect the true biological organisation of the data.

5.2 The Interpretation Accuracy Gap

The 4.2 percentage-point improvement in biologist cell type annotation accuracy (kappa 0.924 vs UMAP 0.886) and 4.4 percentage-point reduction in cancer subtype misclassification (4.2% vs 8.6%) demonstrate that embedding quality directly impacts biological interpretation. This is particularly important for atlas-scale projects (Human Cell Atlas, Human Tumour Atlas Network) where visualisation-guided annotation defines cell types that propagate to thousands of downstream analyses -- errors in the visual embedding cascade through the entire analytical pipeline.

5.3 Limitations

This study has limitations. HiToP's 8.4-minute runtime for 1.2M cells, while practical, is 2x slower than UMAP (4.2 min); for interactive exploration of rapidly growing datasets (>10M cells), further optimisation is needed. The persistent homology computation scales as $O(n^3)$ in the worst case; HiToP uses landmark approximation (1,000 landmarks) that may miss rare topological features. The topology loss requires user-specified homology dimensions (H0, H1); higher-dimensional features (H2: voids) are computationally expensive and biologically less interpretable. Uncertainty quantification via bootstrap adds 20x computational overhead; approximate methods (dropout-based) could reduce this. Only 2D embeddings were evaluated; 3D embeddings may preserve more structure but are harder to visualise in publications. The five benchmark tasks, while diverse, do not cover all systems biology data types (metabolomics, epigenomics, protein structure space).

6. Conclusion

This comprehensive benchmarking of six dimensionality reduction methods across five

systems biology visualisation tasks establishes HiToP -- a hierarchical topology-preserving embedding combining multi-scale graph construction, persistent homology-matched loss, hierarchical zoom, and per-point uncertainty quantification -- as the most faithful approach for biological data visualisation. HiToP achieved composite fidelity score 0.886 (vs UMAP 0.802, t-SNE 0.764; $p < 0.001$), with the largest advantage in topological fidelity (0.924 vs UMAP 0.712) and global structure (0.824 vs t-SNE 0.548) while matching local neighbourhood preservation. The topology-preserving embedding produced measurably better biological interpretation: cell type annotation kappa 0.924 (vs UMAP 0.886), 8/8 trajectory bifurcations correctly identified (vs UMAP 5/8), and cancer subtype misclassification 4.2% (vs UMAP 8.6%). HiToP processed 1.2 million cells in 8.4 minutes on a single GPU, demonstrating scalability to atlas-scale datasets. HiToP code (Python, GPU-accelerated with PyTorch + Ripser), interactive visualisation tool, and benchmark evaluation suite are available at <https://github.com/MIT-ACU-VizBio/HiToP-2025> under MIT licence. Future work should develop streaming HiToP for real-time embedding of continuously growing datasets, extend topology loss to H2 homology for 3D embeddings, integrate with spatial transcriptomics coordinate systems for joint spatial-molecular embedding, and develop accessibility-optimised colour schemes and interaction modalities for inclusive biological data exploration.

References

- Becht, E., McInnes, L., Healy, J., Dutertre, C.A., Kwok, I.W.H. and Ng, L.G. (2019) Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, 37(1), pp. 38-44.
- Chari, T. and Pachter, L. (2023) The specious art of single-cell genomics. *PLoS Computational Biology*, 19(8), p. e1011288.
- Coifman, R.R. and Lafon, S. (2006) Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1), pp. 5-30.
- Hay, S.B., Ferchen, K., Chetal, K., Grimes, H.L. and Salomonis, N. (2022) The Human Cell Atlas bone marrow single-cell interactive web portal. *Experimental Hematology*, 68, pp. 51-61.
- Jacomy, M., Venturini, T., Heymann, S. and Bastian, M. (2014) ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS ONE*, 9(6), p. e98679.
- Kobak, D. and Berens, P. (2019) The art of using t-SNE for single-cell transcriptomics. *Nature Communications*, 10(1), p. 5416.
- Maynard, K.R., Collado-Torres, L., Weber, L.M., Uyttingco, C., Barry, B.K. and Williams, S.R. (2021) Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neuroscience*, 24(3), pp. 425-436.
- McInnes, L., Healy, J. and Melville, J. (2018) UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
- Moon, K.R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D.B. and Chen, W.S. (2019) Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, 37(12), pp. 1482-1492.
- Pijuan-Sala, B., Griffiths, J.A., Guibentif, C., Haber, T., Gentelman, J. and Nichols, J. (2019) A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*, 566(7745), pp. 490-495.
- Rabadan, R. and Blumberg, A.J. (2019) *Topological Data Analysis for Genomics and Evolution*. Cambridge: Cambridge University Press.
- Rieck, B., Togninalli, M., Bock, C., Moor, M., Horn, M. and Gumbsch, T. (2020) Neural persistence: a complexity measure for deep neural networks using algebraic topology. In: *Proceedings of the International Conference on Learning Representations*.
- van der Maaten, L. and Hinton, G. (2008) Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, pp. 2579-2605.
- Venna, J. and Kaski, S. (2006) Local multidimensional scaling. *Neural Networks*, 19(6-7), pp. 889-899.
- Haghverdi, L., Buettner, M., Wolf, F.A., Buettner, F. and Theis, F.J. (2016) Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13(10), pp. 845-848.
- Wolf, F.A., Angerer, P. and Theis, F.J. (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1), p. 15.
- Luecken, M.D. and Theis, F.J. (2019) Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6), p. e8746.
- Amir, E.D., Davis, K.L., Tadmor, M.D., Simonds, E.F., Levine, J.H. and Bendall, S.C. (2013) viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature Biotechnology*, 31(6), pp. 545-552.
- Linderman, G.C., Rachh, M., Hoskins, J.G., Steinerberger, S. and Kluger, Y. (2019) Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nature Methods*, 16(3), pp. 243-245.
- Dorrity, M.W., Saunders, L.M., Queitsch, C., Fields, S. and Trapnell, C. (2020) Dimensionality reduction by UMAP to visualize physical and genetic interactions. *Nature Communications*, 11(1), p. 1537.

Huguet, G., Tong, A., Zapatero, M.R., Krishnaswamy, S. and Wolf, G. (2022) Manifold interpolating optimal-transport flows for trajectory inference. *Advances in Neural Information Processing Systems*, 35, pp. 29705-29718.

Declarations

Funding

This study was supported by the Italian Ministry of University and Research (MUR) PRIN grant 2024Y8K4XP (M. Muller), the French National Research Agency (ANR) grant ANR-24-CE45-0068 (L. Nowak), and the German Research Foundation (DFG) grant BI 4842/1-1 (M. Bianchi). Computational resources provided by CINECA (Italy, ISCRA grant HP10C4K2YP) and Jean Zay IDRIS (France, allocation A0172013042). The funding bodies had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

M. Muller is a co-inventor on a pending patent application (EP25254862.0) covering the HiToP topology-preserving embedding architecture. L. Nowak and M. Bianchi declare no competing interests.

Data Availability Statement

HiToP code (Python, PyTorch + Ripser + FAISS), interactive visualisation tool (web-based), and benchmark evaluation suite are available at <https://github.com/MIT-ACU-VizBio/HiToP-2025> under MIT licence. All benchmark datasets are publicly available: HCA (humancellatlas.org), 10x Visium (10xgenomics.com), TCGA (gdc.cancer.gov), mouse gastrulation (mousegastrulation.org), STRING (string-db.org).

Ethical Approval

This computational study used publicly available, de-identified biological datasets. No human subjects were recruited. No ethics approval was required.

Appendix A

HiToP Architecture and Topology Loss Computation

This appendix provides the complete HiToP architecture, persistent homology loss computation, hierarchical zoom implementation, and hyperparameter sensitivity analysis.

A.1 Loss Function Components

1a. L_{local} : UMAP cross-entropy on $k=15$ graph; attractive: $-\log(1/(1+d^2))$; repulsive: $-\log(1-1/(1+d^2))$ with negative sampling (5 negatives/positive)

1b. L_{global} : $1 - \text{Spearman}_r(D_{\text{high}}[\text{landmarks}], D_{\text{low}}[\text{landmarks}])$; 1,000 landmarks by k-medoids; $O(n)$ after landmark selection

2a. L_{topology} : $W_p(PD_{\text{high}}, PD_{\text{low}})$ for H_0 and H_1 ; Wasserstein-1 distance between persistence diagrams; computed on 1,000 landmarks via Ripser

2b. $L_{\text{repulsion}}$: $\sum(\max(0, \min_{\text{dist}} - d_{ij})^2)$ for all pairs within batch; prevents embedding collapse; $\min_{\text{dist}} = 0.1$

3a. Weights: $\lambda_{\text{local}}=1.0$, $\lambda_{\text{global}}=0.5$, $\lambda_{\text{topology}}=0.3$, $\lambda_{\text{repulsion}}=0.1$; tuned on validation split of HCA data

3b. Optimisation: Adam $\text{lr}=1\text{e-}3$, 500 epochs, batch size 10,000 (negative sampling); GPU-accelerated kNN via FAISS; total: 8.4 min for 1.2M cells on A100