VenusGT: A Trajectory-Aware Graph Transformer for Rare-Cell Discovery in Single-Cell Multi-Omics

Natalia Sikora[1], Rebecca Rees[2], Sean Righardt Holm[3] , Hanchi Ren[*,4], Lewis W. Francis[*,2]
[1]Department of Physics, Swansea University, Swansea SA2 8PP, United Kingdom
[2]Department of Medicine, Swansea University, Swansea SA2 8QA, United Kingdom
[3]London, United Kingdom
[4]Department of Computer Science, Swansea University, Swansea, United Kingdom
Correspondence to: nataliamsikora@gmail.com, L.Francis@swansea.ac.uk

Abstract

Discovering rare subpopulations in single cell multiomics data remains challenging due to their sparse representation and transient nature. Integrating multiple omic modalities often leads to information loss when heterogeneous feature spaces are forced into joint embeddings. We present VenusGT, a trajectory-aware graph transformer pipeline that integrates multimodal single-cell data while preserving biological heterogeneity. VenusGT constructs a heterogeneous cell–gene–peak graph, learns embeddings through attention-based message passing, and incorporates lineage-aware pseudotime information to guide learning. To emphasise transitional dynamics, it applies rarity-weighted, trajectory-guided sampling and a weighted objective that amplifies gradients from rare populations. The model captures temporal continuity through smoothness regularisation and biases attention toward temporally adjacent neighbors. Applied to matched lymphoma dataset, VenusGT improves identification of rare and transitional cell states over existing approaches, enabling interpretable discovery of rare, lineage-specific, and reprogramming cell types in complex single-cell systems.

1 Introduction

The development of modern single-cell (sc) sequencing technologies has provided unprecedented detail of cellular heterogeneity generating deep insights into rare cell populations, and not available in conventional bulk RNA-seq[1].

Rare populations are particularly relevant to tumour initiation, progression, and immune responses. For example, rare stem cell-like memory T cells in premanufacture of CD8+ chimeric antigen receptor T cell therapy proved predictive of outcome for treating patients with diffuse large B-cell lymphoma, highlighting the utility of accurate rare cell identification [2].

Targeting rare subpopulations remains critical for tailored treatment of certain conditions, but can be hindered by limited identification methods, e.g. in acute myeloid leukaemia (AML) quiescent leukaemic stem cells can survive chemotherapy and re-enter the cell cycle, triggering relapse [3, 4].

Learning from one omic modality is often insufficient for identifying very rare cellular subpopulations, as transcriptomic data alone is unable to resolve molecularly similar yet functionally distinct groups [5].

[*]These authors jointly supervised this work.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Multimodal Representation Learning for Healthcare.

Since highly detailed insight into cellular complexity requires multimodal approaches, this necessitates the selection of an appropriate sc data integration framework. In the context of robustness and cellular complexity, different data points related to rare subpopulation are often scarce within modalities and computational tools struggle to maintain both accuracy and efficiency [6], and overcorrect batch effects or suppress rare signals during alignment (Sec. A.1). Joint embeddings of scRNA-seq and scATAC-seq can further obscure modality-specific diversity [7]. Vertical integration of paired data is limited by experimental cost, while horizontal integration across donors introduces batch variability that may mask true rare populations [8]. Early rare-cell discovery methods such as RaceID, GiniClust, CellSIUS, and FiRE prioritised sensitivity but struggled with sparsity and scalability [9]. Graph-based and deep probabilistic models improved multimodal integration. MarsGT [10] applied a graph transformer to scRNA-seq and scATAC-seq data for rare-cell detection but relied on preprocessed inputs and lacked trajectory modelling. scCross [11] identifies cross-sample biclusters robust to batch effects but is restricted to scRNA-seq. Lightweight clustering tools such as scSID [9] improve runtime but assume separable rare clusters, limiting application to continuous trajectories.

Despite progress, existing methods seldom incorporate biological priors such as lineage structure or pseudotime continuity. VenusGT addresses this gap by extending MarsGT with lineage-aware trajectory inference, rarity-aware sampling, and temporally biased attention, enhancing the detection of rare and transitional populations in heterogeneous single-cell graphs.

2 VenusGT: Lineage-Aware Graph Transformer

VenusGT integrates scRNA-seq and scATAC-seq data within a unified heterogeneous graph transformer framework that embeds biological and temporal structure directly into model learning.

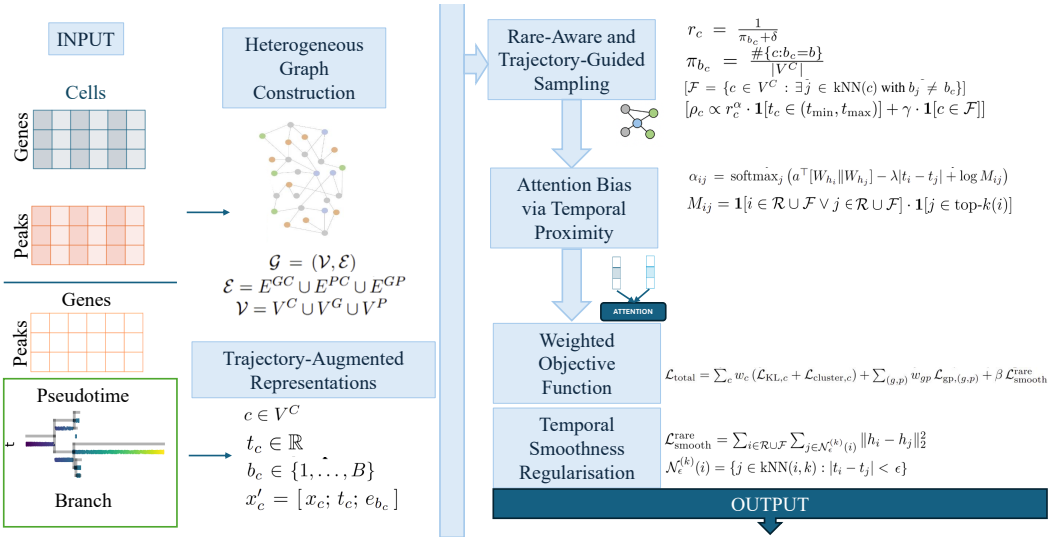

Figure 1: Overview of the VenusGT workflow. The pipeline sequentially constructs a heterogeneous cell–gene–peak graph, augments cell features with pseudotime and lineage information, performs rarity- and trajectory-guided sampling, and learns embeddings through temporally biased attention. Training is guided by a rarity-weighted objective and temporal smoothness regularisation to emphasise rare and transitional cellular states.

The method consists of four core components (Figure 2): (1) heterogeneous graph construction, (2) trajectory-augmented representation, (3) rarity- and trajectory-guided sampling with temporal attention bias, and (4) a rarity-weighted objective with temporal smoothness regularisation.

2.1 Methodology

Heterogeneous Graph Construction: The representation of the input data is formed as a heterogeneous graph with nodes of cells, genes, and peaks. The graph captures their complex interactions

in a unified framework, enabling joint embeddings that reveal cross-modal relationships. This structure also supports effective message passing to improve data integration and reduce issues like dropout in single-cell data. Assuming $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to be a heterogeneous graph with node sets: $\mathcal{V} = V^C \cup V^G \cup V^P$ and $\mathcal{E} = E^{GC} \cup E^{PC} \cup E^{GP}$, where V^C, V^G and V^P are all cells, genes and peaks; E^{GC}, E^{PC} and E^{GP} represent cell-gene, cell-peak, and gene-peak interactions, respectively. Each node $v \in \mathcal{V}$ is initialized with a feature vector $x_v \in \mathbb{R}^{d_v}$.

Data Embedding: We use $h_v^{(\ell)} \in \mathbb{R}^d$ representing the embeddings of each note from the ℓ_{th} layer. The embeddings are learned via stacked multi-head attention layers: $h_v^{(\ell+1)} = \mathrm{GNNLayer}^{(\ell)}\Big(h_v^{(\ell)}, \{h_u^{(\ell)} : u \in \mathcal{N}(v)\}\Big)$, where $\mathcal{N}(v)$ denotes type-aware neighbors of v.

Trajectory-Augmented Representations: Each cell $c \in V^C$ is annotated with pseudotime $t_c \in \mathbb{R}$ and branch label $b_c \in \{1, \ldots, B\}$. We define the augmented feature vector: $x'_c = [\, x_c;\, t_c;\, e_{b_c}\,]$, where e_{b_c} is a one-hot vector encoding the cell's branch. This embeds both temporal progression and lineage identity into the model.

Rare-Aware and Trajectory-Guided Sampling: To prioritize rare and transitional states, we define a rarity score: $r_c = \frac{1}{\pi_{b_c}+\delta}$, where $\pi_{b_c} = \frac{\#\{c:b_c=b\}}{|V^C|}$ denotes the number of cells c whose branch label b_c equals b, and $\delta > 0$ prevents division by zero. We define the frontier set $[\mathcal{F} = \{c \in V^C : \exists\, j \in \mathrm{kNN}(c) \text{ with } b_j \neq b_c\}]$, consisting of cells that have at least one neighbor with a different label. They indicate cells at lineage transitions. Then cells are sampled with $[\rho_c \propto r_c^\alpha \cdot \mathbf{1}[t_c \in (t_{\min}, t_{\max})] + \gamma \cdot \mathbf{1}[c \in \mathcal{F}]]$, focusing on rare ($r_c$) and mid-pseudotime ($t_c$) cells. Neighborhoods are expanded 1-2 hops around these seeds to form subgraphs.

Attention Bias via Temporal Proximity: Attention scores in GNN layers are biased toward temporally adjacent nodes $\alpha_{ij} = \mathrm{softmax}_j\left(a^\top [W_{h_i} \| W_{h_j}] - \lambda |t_i - t_j| + \log M_{ij}\right)$, where λ is weight controlling how strongly temporal distance suppresses attention; $M_{ij} = \mathbf{1}[i \in \mathcal{R} \cup \mathcal{F} \vee j \in \mathcal{R} \cup \mathcal{F}] \cdot \mathbf{1}[j \in \text{top-}k(i)]$ means only edges involving rare or frontier cells are considered and only among the top-k neighbors, reinforcing local structure. This mask restricts attention to edges relevant for rare or transitional dynamics. The attention mechanism is biased to focus on neighbors that are temporally close in pseudotime, and specifically for rare or transition cells. Connections that are not in top-k neighbors or not involving rare or frontier cells are effectively masked out. This ensures that, during message passing, more attention is paid to neighbors that are both close in pseudotime and part of rare or transitional dynamics.

Weighted Objective Function: The final weighted objective function up-weights loss contributions from rare cells and potential rare gene–peak pairs, ensuring the model pays more attention to rare biological signals. Loss terms are reweighted by rarity to prioritize meaningful gradients: $\mathcal{L}_{\text{total}} = \sum_c w_c \left(\mathcal{L}_{\text{KL},c} + \mathcal{L}_{\text{cluster},c}\right) + \sum_{(g,p)} w_{gp}\, \mathcal{L}_{\text{gp},(g,p)} + \beta\, \mathcal{L}_{\text{smooth}}^{\text{rare}}$, where $w_c = r_c^\alpha$ indicates standard rarity-based reweighting. Rare cells get larger weights because r_c is higher when the population is small. α controls how aggressively rarity influences weighting. $\mathcal{L}_{\text{smooth}}^{\text{rare}}$ is temporal smoothness regularisation which is introduced below. $\sum_c w_c \left(\mathcal{L}_{\text{KL},c} + \mathcal{L}_{\text{cluster},c}\right)$ are standard VAE-style KL divergence loss and clustering loss, but scaled to prioritize rare cells. This weighted objective function enables training with smaller batches focused on rare populations without sacrificing gradient fidelity. In other words, rare cells matter more during training.

Temporal Smoothness regularisation: Temporal smoothness is imposed on rare and frontier cells $\mathcal{L}_{\text{smooth}}^{\text{rare}} = \sum_{i \in \mathcal{R} \cup \mathcal{F}} \sum_{j \in \mathcal{N}_\epsilon^{(k)}(i)} \|h_i - h_j\|_2^2$, where $\mathcal{N}_\epsilon^{(k)}(i) = \{j \in \mathrm{kNN}(i,k) : |t_i - t_j| < \epsilon\}$ represents neighbors that are both spatially close (kNN) and temporally nearby (pseudotime constraint), ensuring locally smooth embeddings along pseudotime; $\mathcal{R}$ denotes rare cells. For rare or transitional cells, it ensures their embeddings change smoothly over pseudotime by forcing them to stay close to their temporally adjacent neighbors in latent space.

2.2 Efficiency and Trade-Offs

By restricting sampling, message passing, and regularisation to the subset $\mathcal{R} \cup \mathcal{F}$, VenusGT reduces computation and memory usage, while maintaining rare-cell detection and temporal continuity. This makes it scalable to large single-cell multimodal datasets.

3 Experimental Evaluation

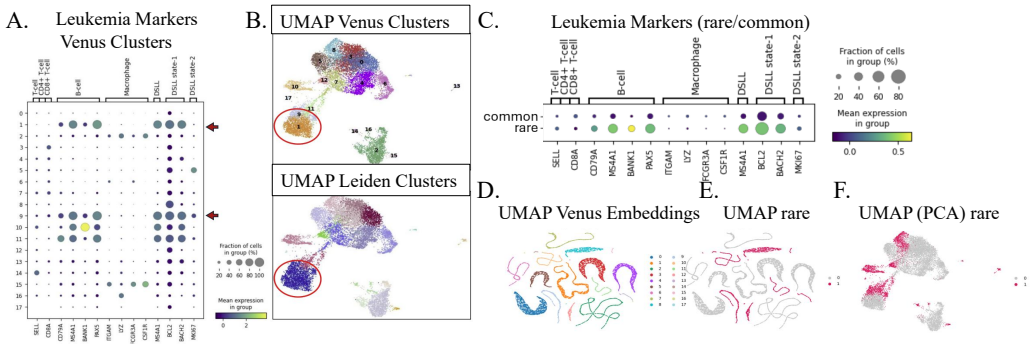

Figure 2: Trajectory-aware graph learning highlights rare and transitional cellular states. (A) UMAP embedding highlighting rare-cell candidates identified by VenusGT's rarity-aware sampling. (B) Latent representations learned by the VenusGT reveal 63 discrete cellular clusters capturing fine-grained transcriptional states. (C) and (D) Lineage milestones, pseudotime and branch ordering utilised in trajectory-augmented representations in the augmented feature vector.

We evaluated VenusGT on a matched scRNA-seq and scATAC-seq dataset from a patient with diffuse small lymphocytic lymphoma (DSLL) (Sec. C.1). MarsGT identified four B-cell clusters, including one normal and three malignant subtypes: B lymphoma-state-1 (BLS1), B lymphoma-state-2 (BLS2), and B lymphoma-state-3 (BLS3). BLS1 is the rare population. VenusGT reproduced these clusters but additionally resolved subclusters within malignant groups, revealing finer-grained heterogeneity and transitional states.

When comparing clustering results of VenusGT to commonly used Leiden clustering, VenusGT was able to resolve the B cell clusters 1 and 9 which were represented as a single cluster by Leiden clustering (Figure 3). Cluster 9 was identified by marker gene expression as the rare cell state presenting DSLL state-2.

Temporal ordering of embeddings aligned with lineage milestones and pseudotime progression, validating that the model preserves developmental continuity while enhancing rare states sensitivity.

4 Discussion

VenusGT extends multimodal graph transformers by embedding lineage and temporal priors into graph representation learning. Unlike MarsGT, which operates on static batch-corrected embeddings, VenusGT dynamically integrates pseudotime and branch identity to guide attention, sampling, and optimisation. The rarity-weighted objective ensures that rare cell states contribute proportionally more to learning, mitigating imbalance inherent in biological data. Temporal smoothness and attention bias further stabilise embeddings, maintaining both global structure and local trajectory coherence.

We reduced computational cost (Sec. D) by limiting message passing and regularisation to informative subsets (rare and frontier cells), achieving scalability without sacrificing detection accuracy. Conceptually, VenusGT bridges multimodal graph learning with trajectory inference, forming a biologically informed manifold representation suited to both rare-cell discovery and lineage reconstruction.

5 Impact and relevance

VenusGT introduces a principled framework for dynamic single-cell representation learning that couples heterogeneous graph modeling with pseudotemporal inference. Methodologically, it generalises transformer-based graph learning to dynamic manifolds, enabling interpretable, temporally coherent embeddings under multimodal sparsity. Beyond single-cell analysis, this paradigm could be readjusted to other temporal graph systems, such as disease progression modeling, developmental dynamics, or evolving biological networks, where entities interact through heterogeneous relationships that evolve over time.

References

[1] Getnet Molla Desta and Alemayehu Godana Birhanu. Advancements in single-cell rna sequencing and spatial transcriptomics: transforming biomedical research. *Acta Biochimica Polonica*, 72, February 2025. ISSN 1734-154X. doi: 10.3389/abp.2025.13922. URL http://dx.doi.org/10.3389/abp.2025.13922.

[2] Yao Wang, Chuan Tong, Yuting Lu, Zhiqiang Wu, Yelei Guo, Yang Liu, Jianshu Wei, Chunmeng Wang, Qingming Yang, and Weidong Han. Characteristics of premanufacture CD8+T cells determine CAR-T efficacy in patients with diffuse large b-cell lymphoma. *Signal Transduction and Targeted Therapy*, 8(1):409, October 2023. ISSN 2059-3635. doi: 10.1038/s41392-023-01659-2. URL http://dx.doi.org/10.1038/s41392-023-01659-2.

[3] Sophie G Kellaway, Sandeep Potluri, Peter Keane, Helen J Blair, Luke Ames, Alice Worker, Paulynn S. Chin, Anetta Ptasinska, Polina K. Derevyanko, Assunta Adamo, Daniel J. L. Coleman, Naeem Khan, Salam A. Assi, Anja Krippner-Heidenreich, Manoj Raghavan, Peter N. Cockerill, Olaf Heidenreich, and Constanze Bonifer. Leukemic stem cells activate lineage inappropriate signalling pathways to promote their growth. *Nature Communications*, 15(1): 1359, February 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-45691-4. URL http://dx.doi.org/10.1038/s41467-024-45691-4.

[4] Amir Hosseini, Abhinav Dhall, Nemo Ikonen, Natalia Sikora, Sylvain Nguyen, Yuqi Shen, Maria Luisa Jurgensen Amaral, Alan Jiao, Felice Wallner, Philipp Sergeev, Yuhua Lim, Yuanqin Yang, Binje Vick, Kimihito Cojin Kawabata, Ari Melnick, Paresh Vyas, Bing Ren, Irmela Jeremias, Bethan Psaila, Caroline A. Heckman, M. Andrés Blanco, and Yang Shi. Perturbing lsd1 and wnt rewires transcription to synergistically induce aml differentiation. *Nature*, 642 (8067):508–518, April 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-08915-1. URL http://dx.doi.org/10.1038/s41586-025-08915-1.

[5] Jongsu Lim, Chanho Park, Minjae Kim, Hyukhee Kim, Junil Kim, and Dong-Sung Lee. Advances in single-cell omics and multiomics for high-resolution molecular profiling. *Experimental and Molecular Medicine*, 56(3):515–526, March 2024. ISSN 2092-6413. doi: 10.1038/s12276-024-01186-2. URL http://dx.doi.org/10.1038/s12276-024-01186-2.

[6] Mingguang Shi and Xuefeng Li. Addressing scalability and managing sparsity and dropout events in single-cell representation identification with zigacl. *Briefings in Bioinformatics*, 26(1), November 2024. ISSN 1477-4054. doi: 10.1093/bib/bbae703. URL http://dx.doi.org/10.1093/bib/bbae703.

[7] Zhenchao Tang, Jiehui Huang, Guanxing Chen, and Calvin Yu-Chian Chen. Comprehensive view embedding learning for single-cell multimodal integration. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(14):15292–15300, March 2024. ISSN 2159-5399. doi: 10.1609/aaai.v38i14.29453. URL http://dx.doi.org/10.1609/aaai.v38i14.29453.

[8] Malte D Luecken, M Büttner, K Chaichoompu, A Danese, M Interlandi, M F Mueller, D C Strobl, L Zappia, M Dugas, M Colomé-Tatché, and Fabian J Theis. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1):41–50, January 2022. ISSN 1548-7105. doi: 10.1038/s41592-021-01336-8. URL http://dx.doi.org/10.1038/s41592-021-01336-8.

[9] Shudong Wang, Hengxiao Li, Kuijie Zhang, Hao Wu, Shanchen Pang, Wenhao Wu, Lan Ye, Jionglong Su, and Yulin Zhang. scsid: A lightweight algorithm for identifying rare cell types by capturing differential expression from single-cell sequencing data. *Computational and Structural Biotechnology Journal*, 23:589–600, December 2024. ISSN 2001-0370. doi: 10.1016/j.csbj.2023.12.043. URL http://dx.doi.org/10.1016/j.csbj.2023.12.043.

[10] Xiaoying Wang, Maoteng Duan, Jingxian Li, Anjun Ma, Gang Xin, Dong Xu, Zihai Li, Bingqiang Liu, and Qin Ma. MarsGT: Multi-omics analysis for rare population inference using single-cell graph transformer. *Nature Communications*, 15(1):338, January 2024. ISSN 2041-1723. doi: 10.1038/s41467-023-44570-8. URL http://dx.doi.org/10.1038/s41467-023-44570-8.

[11] Alexander Gerniers, Siegfried Nijssen, and Pierre Dupont. sccross: efficient search for rare subpopulations across multiple single-cell samples. *Bioinformatics*, 40(6), June 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae371. URL `http://dx.doi.org/10.1093/bioinformatics/btae371`.

Author Contributions

N.S. designed the methodology and implemented the VenusGT model, performed data analysis, wrote the manuscript, and created figures. R.R. contributed biological interpretation and contributed towards figure generation and the manuscript write-up. S.R.H. provided methodological input, biological interpretation, advised on bioinformatics workflows, and helped define markers for rare cellular subpopulations. H.R. contributed algorithmic refinements and describing the framework in the methodology section, as well as supervised the machine learning components of the project, and provided feedback on the final manuscript. L.W.F. provided biological supervision, interpretation of results, and overall project oversight.

A Paper motivation

Cellular differentiation and disease progression often unfold through continuous, branching trajectories, where rare and transient cell states carry disproportionate biological importance, including playing a role as treatment resistant precursors, reprogramming intermediates, or early malignant transformations. Traditional sc analysis methods struggle to detect these states because they are sparsely represented, often obscured by noise, and easily lost when multimodal data (e.g., RNA and chromatin accessibility) are forced into joint embeddings. At the same time, pseudotime relationships and lineage structure, which are central to understanding how cell states evolve, are usually ignored during multimodal integration, leading to embeddings that preserve global clusters but miss dynamic biological transitions. VenusGT addresses this gap, enabling the discovery of rare and transitional populations while preserving the biological continuity of developmental lineages.

A.1 Modal gap

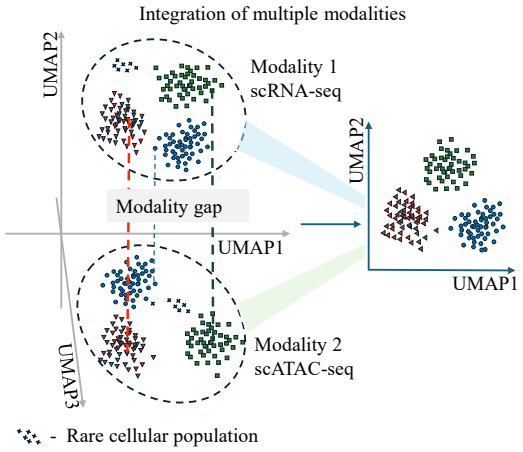

Figure 3: The multimodal integration aims to reduce the modality gaps whilst preserving biological heterogeneity. Unfortunately, many algorithms end up removing rare cells, treating them as noise.

B Limitations

Although VenusGT improves rare-cell discovery by integrating multimodal signals with trajectory-aware graph learning, several limitations should be acknowledged. First, the framework needs to be validated across multiple datasets and the model relies on pseudotime and branch labels computed

using diffusion pseudotime. These labels depend on the quality of the RNA derived manifold and may be sensitive to preprocessing choices, root cell selection, or branching resolution. As a consequence, trajectory structure is imposed rather than jointly inferred, and inaccuracies in pseudotime may propagate into the attention mechanism and loss weighting.

Second, our evaluation uses a single patient-derived lymph node sample profiled with a specific 10x Genomics Multiome chemistry. Although paired scRNA-seq and scATAC-seq data enable multimodal integration, the dataset is restricted to one tissue, one disease context, and one acquisition platform. Broader benchmarking across donors, tissues, disease states, and sequencing platforms is required to assess generalisability and robustness to biological and technical variability.

Finally, although VenusGT reduces computational burden by restricting temporal masking, sampling, and regularisation to rare and frontier subsets, the approach still requires computing kNN graphs, diffusion maps, and large heterogeneous graph embeddings. Scaling to atlas-level datasets or higher modality counts may require additional optimisations or approximate neighborhood search.

Overall, VenusGT provides a principled framework for rare cell discovery, but future work should incorporate cross dataset evaluation, and expanded regulatory modeling.

C Experimental setup

C.1 Dataset

Dataset. We analyzed a paired scRNA–seq and scATAC–seq dataset which we downloaded from 10x genomics platform. The dataset is available at: Dataset link (10x Genomics). It was generated from a flash-frozen intra-abdominal lymph node tumour obtained from a patient with diffuse small lymphocytic lymphoma (BioIVT Asterand®).

Paired scATAC-seq and gene expression libraries were prepared using the *Chromium Next GEM Single Cell Multiome ATAC + Gene Expression* chemistry (CG000338 Rev A). Libraries were sequenced on an Illumina NovaSeq 6000 v1.5 instrument using the reverse-complement dual-index workflow. Nuclei loading targeted a recovery of approximately 15,000 nuclei to account for potential overcounting following FACS cleanup.

The final dataset contained an estimated 14,566 nuclei. For the scATAC modality, the data comprised a mean of 32,248 raw read pairs per cell, a median of 8,396 high-quality fragments per cell, and 109,789 called peaks. For the gene expression modality, nuclei had a mean of 25,917 raw reads per cell, a median of 1,186 detected genes, and a median of 1,671 UMI counts per cell. Across modalities, 12,382 genes were linked to 58,474 peaks.

The dataset is publicly released by 10x Genomics under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

C.2 Methodology

Pseudotime inference. Pseudotime for VenusGT was derived using diffusion pseudotime (DPT) implemented in Scanpy, following standard preprocessing steps for scRNA–seq data. Raw counts were normalised to 10^4 reads per cell, log-transformed, and the top 2,000 highly variable genes were selected. Gene expression values were scaled to unit variance (clipped at 10) and a principal component analysis was performed using up to 50 components.

We first computed a global DPT pseudotime by selecting a biologically early root cell based on canonical early B-cell markers and its position along the leading diffusion component. This provides a smooth one-dimensional ordering along the dominant trajectory.

Since DPT alone does not infer branching structure, we used Partition-based Graph Abstraction (PAGA) to identify the coarse-grained topology of the transcriptomic manifold. Branch labels were assigned to cells based on their most likely PAGA lineage paths, and pseudotime was used as an additional continuous descriptor within each branch.

D Other

MarsGT took over 4h on the same dataset.

Table 1: Computational cost and graph statistics for the VenusGT pipeline.

Wall-clock runtime	
Matrix loading (Step 01)	116.50 s
Node feature SVDs (Step 02)	203.11 s
Scanpy pseudotime (Step 03)	0.42 s
kNN graph construction (Step 04)	4.81 s
Training loop (100 steps, Step 05)	5.40 s
Graph size	
Number of cells (N_{cells})	14,148
Number of genes (N_{genes})	19,107
Number of peaks (N_{peaks})	109,789
Cell–gene edges	21,402,192
Cell–peak edges	65,122,974