CELLMEMORY: HIERARCHICAL INTERPRETATION OF OUT-OF-DISTRIBUTION CELLS USING BOTTLENECKED TRANSFORMER

Qifei Wang[†,1], **He Zhu**[†,2], **Yiwen Hu**[1], **Yanjie Chen**[1], **Yuwei Wang**[3], **Xuegong Zhang**[4], **James Zou**[5], **Manolis Kellis**[6], **Yue Li**[*,2], **Dianbo Liu**[*,7], **Lan Jiang**[*,1]

[1]China National Center for Bioinformation, CAS [2]McGill University
[3]Institute of Psychology, CAS [4]Tsinghua University [5]Stanford University
[6]Massachusetts Institute of Technology [7]National University of Singapore

ABSTRACT

Applying machine learning to cellular data presents several challenges. One such challenge is making the methods interpretable concerning both the cellular information and its context. Another less-explored challenge is the accurate representation of cells outside existing references, referred to as out-of-distribution (OOD) cells. OOD cells arise from physiological conditions (e.g., diseased vs. healthy) or technical variations (e.g., single-cell references vs. spatial queries). Inspired by the Global Workspace Theory in cognitive neuroscience, we introduce CellMemory, a bottlenecked Transformer with improved generalization designed for the hierarchical interpretation of OOD cells. CellMemory outperforms large-scale foundation models pre-trained on tens of millions of cells, even without pre-training. Moreover, it robustly characterizes malignant cells and their founder cells across different patients, revealing cellular changes caused by the diseases. We further propose leveraging CellMemory's capacity to integrate multi-modalities and phenotypic information, advancing toward the construction of VIRTUAL ORGAN.

1 INTRODUCTION

Single-cell sequencing has revolutionized genomics by providing unprecedented insights into cellular heterogeneity and dynamics. It is worth noting that cells from different individuals, technologies, or species do not exhibit similar distribution patterns. For example, there are variations in the states of malignant and healthy cells, as well as heterogeneity among malignant cells across different patients. Moreover, significant technical variations are present between single-cell and spatial omics data. Cells that deviate substantially from the established paradigms are defined as out-of-distribution cells Lotfollahi et al. (2024).

Deciphering OOD cells requires concurrent identity inference and data integration, a process known as reference mapping. Currently, only a very limited number of methods can do both simultaneously De Donno et al. (2023); Hao et al. (2021); Lotfollahi et al. (2022). Furthermore, these methods often fall short of adequately considering gene interactions, which are crucial for generalizing cell representations in population-scale (data size), granule-level (annotation of high-resolution) single-cell atlases. For biological researchers, the insights provided by explainable artificial intelligence (xAI) frequently surpass the value of predictions, as they may enable novel understandings of life processes Novakovsky et al. (2023). The application of xAI to cells outside the reference scope is an area that has not been thoroughly explored Ma et al. (2024).

Generative Pre-trained Transformers (GPT), capable of understanding the interrelationships among features, present a novel strategy for exploring data across various fields. However, current single-cell foundation models struggle with handling long token inputs due to the computational complexity of self-attention mechanisms Cui et al. (2024); Theodoris et al. (2023); Rosen et al. (2023). This limits their capacity to comprehend the regulatory relationships among genes and to elucidate comprehensive cell representations. Moreover, owing to inherent constraints in their frameworks, these Transformer models may fail to provide a high-resolution interpretation of their behavior. In this

study, we developed a neuroscience-inspired Transformer architecture to specifically solve these problems.

The Global Workspace Theory (GWT) in neuroscience suggests that consciousness emerges from the selective dissemination of information via a shared "global workspace" or "memory space" Baars (1988). This workspace, a network of densely interconnected neurons, facilitates competition between neuron modules to write information to this limited-capacity space Goyal et al. (2021). Inspired by this concept, we propose that combining GWT with a Transformer architecture can efficiently manage informative and sparse single-cell data. We have developed CellMemory, which is a bottlenecked architecture within the Transformer that uses a cross-attention mechanism. This unique design allows CellMemory to learn generalized representations from standardized paradigms and perform interpretable inferences for OOD cells. By incorporating various specialist modules into the global workspace transparently, CellMemory allows for a hierarchical interpretation of the model behavior. The simulation of the consciousness formation process in the global workspace enhances our understanding of how the deep learning model systematically processes biological information.

A comprehensive benchmarking study compared 3 single-cell foundation models and 15 task-specific methods to CellMemory. CellMemory demonstrated harmonious integration and accurate label transfer, even outperforming single-cell foundation models that were pre-trained with tens of millions cells regarding generalization ability and computational efficiency Rebecca et al. (2023). Finally, we used CellMemory and healthy references to delineate OOD malignant cells. CellMemory helps us contextualize malignant cells in harmonious embeddings and explain their origins. It underscores the importance of leveraging xAI and existing references to better understand complex diseases by offering valuable insights into OOD disease states.

2 RESULTS

2.1 THE INSPIRATION FOR CELLMEMORY

The Global Workspace Theory in neuroscience suggests that consciousness arises from the selective dissemination of information through a shared "global workspace". The brain consists of numerous specialized modules that handle various tasks like vision, hearing, memory, and emotions. The "global workspace" or "Memory space" is a network of densely interconnected neurons that enables these specialized modules to share and communicate information. Due to its limited capacity, not all information can be retained in the global workspace. Specialized modules compete to write information to the global workspace, which then becomes part of conscious awareness. Once stored, the information is broadcasted to diverse specialized neural modules via synchronized neural firing and rapid signaling. This ensures that various brain regions process the information concurrently, contributing to a unified conscious experience.

Inspired by GWT in cognitive neuroscience, we propose that an architecture comprising specialists trained to communicate effectively through the constraint of a global workspace yields a substantial enhancement in computational efficiency and generalization. To achieve this, we have developed CellMemory, which utilizes a bottlenecked architecture within a Transformer-based framework (Figure 1). CellMemory is tailored to derive meaningful cellular representations from single-cell references and effectively address the challenge of making inferences about OOD cells in an explainable manner. Furthermore, it can integrate population-scale single-cell datasets at a granule level, adeptly managing variations in cell state, technical platform, spatiotemporal conditions, ethnicity, and species.

The conventional self-attention mechanism used by BERT has a quadratic complexity concerning sequence length, resulting in increased time and memory complexity as sequences become larger Duman Keles (2022). This presents a particularly difficult challenge when dealing with large-scale single-cell omics data. In contrast, CellMemory uses the "Specialists" module (length=M) to perform cross-attention with the "Memory" (length=H, $H \ll M$). In this process, a significant amount of biological information competes for the limited "Memory space". The refined information is then broadcast to all modules by the "Memory". The bottleneck of "Memory space" acts as a filter, prioritizing the most significant information and ensuring that the most important biological details are

efficiently communicated. As a result, the CellMemory architecture has impressive generalization capabilities and substantially reduces computational costs.

During the inference process, CellMemory empowers the allocation of different weights to various tokens, or genes, within the input based on their determined significance through attention scores (Interpretation level 1). Simultaneously, it ensures the consideration of the cellular context when integrating all scores into memory slots (memory scores, Interpretation level 2). This unique architecture equips the model to provide interpretability at both gene and gene program levels (Figure 1), thus enabling a more insightful understanding of specific cellular states.

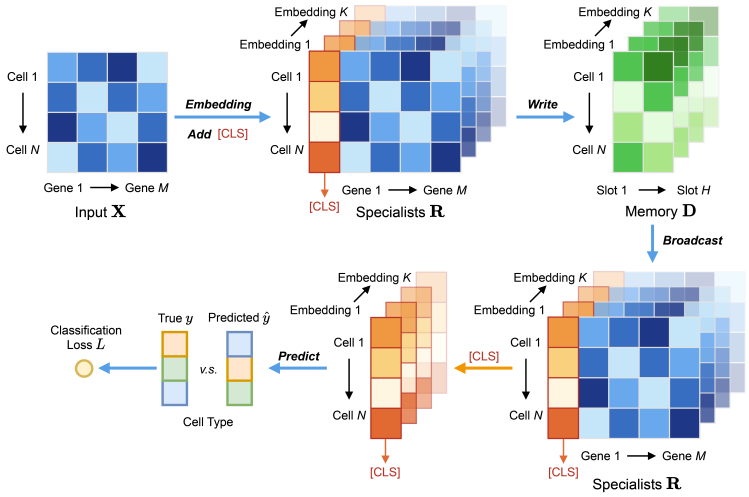

Figure 1: The illustration of CellMemory. CellMemory learns cell representations from references in a supervised manner. The densely interconnected neurons in CellMemory enable specialized modules to share and communicate information. The "Specialists" (specialized modules) compete with each other to write information to the "Memory". The information is then broadcasted to "Specialists" through synchronized neural firing and rapid signaling. The processes of writing and broadcasting between "Specialists" and "Memory" are executed by the cross-attention mechanism. The CLS token of CellMemory is used for training the model by comparing predictions with the real labels.

2.2 CellMemory enables one-stop reference mapping of single-cell atlas across diverse scenarios

We introduce CellMemory, a tool that streamlines reference mapping of single-cell atlases across diverse biological and technological conditions. Its annotation performance was benchmarked against task-specific methods De Donno et al. (2023); Hao et al. (2021); Cui et al. (2024); Theodoris et al. (2023); Rosen et al. (2023); Chen et al. (2023); Domínguez Conde et al. (2022) using 7 datasets exhibiting varied attributes Domínguez Conde et al. (2022); Fasolino et al. (2022); Hrovatin et al. (2023); Jorstad et al. (2023); Kumar et al. (2023); Salcher et al. (2022); Steuernagel et al. (2022) (Figure 2a). Trained on cells from one condition, CellMemory accurately predicted outcomes from distinct states – varying by platform, organ, or species – using metrics such as F1-score and accuracy. The F1-score, in particular, provided a nuanced evaluation for rare cell types in complex single-cell datasets. Notably, CellMemory outperformed single-cell foundation models in multiple tasks, including the identification of rare cell types on most benchmark datasets Tianyu et al. (2023), without the need for resource-intensive pre-training. Although pre-trained foundation models can achieve high accuracy on small-scale datasets, an imbalance in cell proportions during the pre-training phase leads to the neglect of low-abundance cell types. In contrast, CellMemory demonstrated robust generalization and processed data more efficiently than self-attention-based Transformers (Figure 2b).

We further evaluated CellMemory's ability to integrate cross-state and cross-platform datasets by generating cell embeddings (using the same CLS embedding as in the annotation tasks) and comparing its performance with advanced integration tools Luecken et al. (2022) such as scVI Lopez et al. (2018), scGPT, and scPoli (Figure 2c). Benchmark results revealed that CellMemory not only

preserves biological representations but also minimizes batch effects. These outcomes can be attributed to the cross-attention mechanism that is designed to interact with the bottlenecked space and its supervised training approach. Collectively, these strategies enable the model to acquire robust cell representations for downstream analysis.

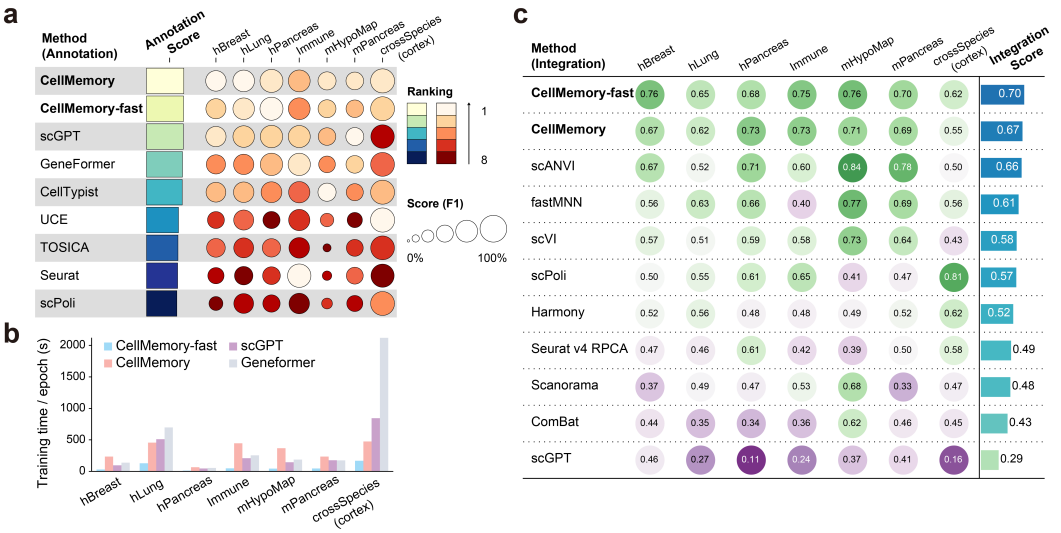

Figure 2: Annotation and integration benchmark. **a**, F1-score (macro) is employed to assess the annotation performance of single-cell datasets. Each circle represents the mean of the five-fold cross-validation for the respective method (except for crossSpecies, which includes data from five species). The overall performance across all datasets corresponds to the Annotation Score. CellMemory: The model was trained using all input genes, considering genes with zero expression in each cell. CellMemory-fast: Genes with zero expression in each cell were filtered out, and the model was trained using the Mixed Precision Training strategy. **b**, The average time taken to train one epoch is recorded for each dataset for CellMemory, scGPT, and Geneformer. **c**, The integration benchmark for the dataset includes popular integration methods such scPoli, scVI, and scGPT. The Integration Score is comprehensively evaluated based on the retention of biological information and the removal of batch effects. The overall score is calculated as a weighted average of the batch correction score and the biological conservation score, with weights of 0.4 and 0.6, respectively.

2.3 CellMemory facilitates the interpretable characterization of single-cell spatial omics

CellMemory facilitates the interpretable characterization of single-cell spatial omics. Deciphering spatial omics is challenging due to significant technical bias and variable information richness compared to standard single-cell data. To address these challenges in characterizing OOD cells, we applied CellMemory's robust representation capabilities to spatial transcriptomics. We benchmarked its spatial annotation performance using data from CosMx Salcher et al. (2022); He et al. (2022), MERFISH Steuernagel et al. (2022); Moffitt et al. (2018), and Slide-seq Chen et al. (2021), comparing it against various spatial analysis methods Aliee & Theis (2021); Biancalani et al. (2021); Cable et al. (2022); Mages et al. (2023); Rodriques et al. (2019) and two single-cell foundation models. Accuracy was used as the primary evaluation metric given the complexity of spatial data, and results demonstrated that CellMemory outperformed all tested methods, even without spatial coordinate information (Figure 3a). Although pre-trained models like scGPT and Geneformer may improve cross-species spatial data comprehension, CellMemory's performance remained superior.

We further investigated granule-level representation of spatial omics using human cortex Silde-tags data Russell et al. (2024). Initially, CellMemory leveraged a single-cell cortex reference from normal samples Mariano et al. (2023) to build an L2-level model (24 cell types). As shown in Figure 3b, it produced appropriate embeddings and advanced annotations for the Slide-tags data, integrating the

distinct omics datasets effectively. Next, a granule-level model (L3, 131 cell types) was constructed by integrating single-cell and Slide-tags data using CLS embeddings (Figure 3c). Integration performance assessed using Slide-tags and hierarchical L1/2/3 annotations revealed coherent, granule-level characterizations (Figure 3d). When examination of CellMemory's interpretation of these OOD cells revealed that although VWC2L is expressed across several clusters, the model uniquely assigns it higher weights in L4 IT_2 cells after considering all gene correlations, indicating that this gene is crucial for the model's inference of these cells. Remarkably, CellMemory can accurately characterize rare cell states, such as Micro-PVM_1 (F13A1+), which exhibit a very low proportion in the reference set (only 0.43‰) and in the spatial data (3.4‰), providing reasonable explanations for the integration result (Figure 3e). This is essential for dissecting spatial cellular heterogeneity in complex biological systems and will significantly enhance our perception of spatial cell states.

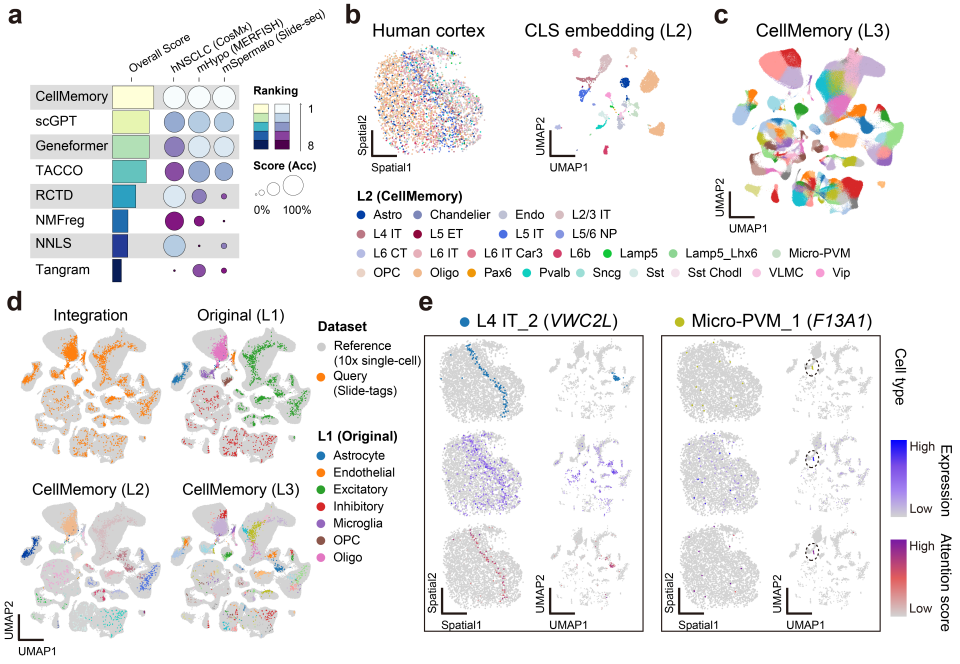

Figure 3: CellMemory interprets single-cell spatial data across omics. **a**, The accuracy of spatial annotation tools is evaluated using datasets from mHypo (mouse hypothalamus measured by MERFISH), hNSCLC (human NSCLC measured by CosMx), and msSermato (mouse spermatogenesis measured by Slide-seq). Each dataset comprised three samples for replication. **b**, CellMemory was trained using single-cell data at the L2 cell type resolution, to generate CLS embeddings and annotations for Slide-tags cells. The right part is plotted by the L1 (original) and L2 (CellMemory) cell types in spatial coordinates. **c**, CellMemory model was built using single-cell data at the L3 resolution, to integrate single-cell and Slide-tags data. **d**, The cells highlighted in the co-embedding are Slide-tags cells, labeled with Slide-tags cell, L1 (original), L2 (CellMemory), and L3 (CellMemory) cell identity. **e**, The identification of Slide-tags cells at L3 resolution by CellMemory (L4 IT_2 and Micro-PVM_1) is displayed (the first line), along with the expression (the second line) and attention score (the third line) of TAGs (VWC2L, F13A1). The left half represents the spatial coordinate, and the right half is the UMAP coordinate. TAGs: Top Attention Genes.

2.4 Studying malignant cells and the heterogeneity among patients using normal cells as a baseline

While it's feasible to map data sets from healthy individuals to healthy reference atlases, interpreting diseased samples using healthy atlases is highly desirable but comes with distinct challenges Dann et al. (2023). Malignant cells are notably intricate, exhibiting aneuploidy and substantial heterogeneity among patients, thereby amplifying the inherent challenges of the task at hand. With CellMemory, we address this issue by utilizing its strong generalization capability, which allows the characterization of malignant cell states using an established healthy reference.

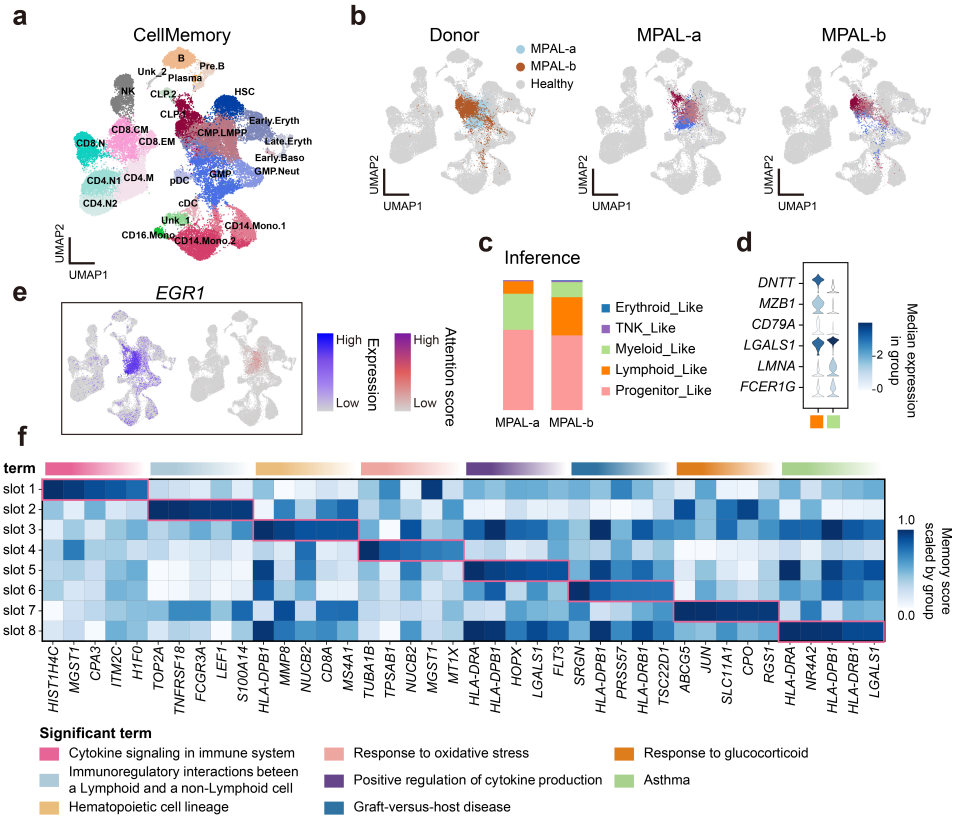

Figure 4: CellMemory interprets the individual heterogeneity of complex diseases. **a**, CellMemory integrated malignant blood cells from MPAL patients with bone marrow and peripheral blood mononuclear cells from healthy samples. **b**, Colored points show the positions of MPAL patient cells (MPAL-a: cells; MPAL-b: cells) in the co-embedding space and their inferred cell identities. **c**, Classify the lineage proportions based on cell types. **d**, The expression of lineage-specific genes in lymphoid and myeloid cells is inferred from MPAL-a patient cells. **e**, Expression and attention score of EGR1 in healthy cells and MPAL-a cells. **f**, The heatmap showing the attention level of progenitor-like cells from MPAL-a patient in memory space. Eight memory slots are displayed, each showing the top five TAGs that it focuses on. The top 1-50 TAGs for each slot were subjected to enrichment analysis, with significant terms displayed above the heatmap.

We focus on Mixed Phenotype Acute Leukemia (MPAL), a high-risk subtype of leukemia characterized by the co-expression of mixed myeloid and lymphoid features. Recent reports have indicated that this high-risk subtype is associated with abnormalities in primitive hematopoietic progenitors Alexander et al. (2018). Accurately characterizing the malignant cell components and molecular signatures in MPAL patients is crucial for studying the potential mechanisms driving this complex disease. In our investigation, we trained CellMemory using bone marrow and peripheral blood mononuclear cells (BMMC, PBMC) from healthy individuals. We then applied this model to analyze malignant cells from two MPAL patients Granja et al. (2019), one with B myeloid MPAL (MPAL-a) and the other with T-myeloid MPAL (MPAL-b) (Figure 4a). CellMemory was utilized to define the lineage proportions in these patients, identifying the progenitor-like state as the predominant original type (Figure 4b).

After examining the specific features of lymphoid and myeloid cells in MPAL-a, we consistently observed characteristics related to the normal lineage and developmental stage of these malignant cells (Figure 4c-d). The gene level (Interpretation L2) and gene program level (Interpretation L2) information stored in the bottlenecked memory space provide biologically meaningful interpretations for these malignant cells with progenitor-like properties. L1: EGR1, a key feature of the progenitor lineage van Galen et al. (2019), was identified as a TAG by CellMemory, highlighting the origin

characteristics of MPAL (Figure 4e). L2: The GO term "Hematopoietic cell lineage" enriched in memory slot 3's top attention genes suggests the cellular origin, while the term "Response to oxidative stress" in slot 4 reflects the cells' stress state Chen et al. (2020). The term in slot 6 indicates immune function dysregulation Negrin (2015) (Figure 4f). Furthermore, CellMemory provided distinct hierarchical interpretations for MPAL-b, with increased attention to DNTT, implying potential differences in lineage or stages for the cell of origin among patients.

The precise characterization of OOD cells with CellMemory reveals the diverse nature of complex tumor origins. These clear and interpretable representations form a robust foundation for understanding the relationship between normal and disease states, underscoring the potential of CellMemory as an invaluable tool for identifying tumor founder cells.

3 DISCUSSION

To take advantage of existing reference data for understanding out-of-distribution (OOD) cells, we have developed CellMemory. CellMemory is a Transformer model that employs a cross-attention mechanism inspired by a theory of consciousness. We conducted a comprehensive comparison of 3 single-cell foundation models and 15 task-specific methods for integration and annotation cells. Due to its generalized representations, CellMemory outperforms other methods sanking scGPT and Geneformer, especially in cross-omics situations. We believe that CellMemory's exceptional performance is due to its bottlenecked architecture's attention and competitive mechanism, which allows it to retain essential information.

We then focused on the comparison of disease versus healthy states to explore interpretable inference for OOD cells. Notably, by analyzing the most focused genes in cell type specific memory slots, we identified significant biological function differences within the gene programs that are targeted by different memory slots. A variety of specialist modules were organized into distinct shared global workspaces within a bottleneck, ensuring consistency in task completion. This process is analogous to the brain coordinating and unifying driving-related visual and auditory information in the consciousness space for safe driving. This hierarchical interpretation facilitates the analysis of OOD objects related to cellular states, further enhancing our understanding of the dynamic manner in which the model processes biological information.

Toward building the VIRTUAL ORGAN. Recent discussions on building the Virtual Cell model have attracted significant attention Bunne et al. (2024). It is widely believed that developing Virtual Cell will substantially enhance our understanding of biological systems and accelerate experimental cycles. However, current single-cell data resources are overwhelmingly dominated by transcriptomics modalities, and the Virtual Cell model based solely on unimodal may not fully capture the complexity of cellular characteristics. Therefore, relying exclusively on hundreds of millions of transcriptomic profiles may be insufficient for constructing a comprehensive Virtual Cell. Interestingly, recent studies indicate that increases in data size and diversity during the pre-training stage offer limited benefits for single-cell foundation model DenAdel et al. (2024). Moreover, as noted in other reports, pre-training strategies developed for Natural Language Models may lack a "foundational" understanding of genomics Vishniakov et al. (2024). These challenges underscore the urgent need to develop more efficient and generalizable computational frameworks capable of handling diverse multimodal data, as well as to explore model training strategies that better align with biological information.

Looking ahead, recent report on massive-scale and cost-effective single-cell multiome sequencing technology indicate that we can be employed to focus on some organs or tissues, generating population-scale multiome data Li et al. (2025). This approach provides unprecedented data richness that deepens our understanding of cellular regulatory processes. For example, soon, by integrating perturbation and organoid resouces, researchers may be able to assess donors status using a Virtual Immune System and simulate readouts across multi-omics, helping to identify reliable intervention targets. To this end, we envision leveraging highly efficient and generalizable computational frameworks, such as the Bottlenecked Transformer, to decipher population-scale single-cell multi-omics data along with their extensive phenotypic information, enabling to construct the Virtual Organ and, subsequently, the Virtual Cell.

REFERENCES

T. B. Alexander, Z. Gu, I. Iacobucci, K. Dickerson, J. K. Choi, B. Xu, and et al. The genetic basis and cell of origin of mixed phenotype acute leukaemia. *Nature*, 562(7727):373–379, 2018. ISSN 0028-0836 (Print) 0028-0836. doi: 10.1038/s41586-018-0436-0.

H. Aliee and F. J. Theis. Autogenes: Automatic gene selection using multi-objective optimization for rna-seq deconvolution. *Cell Syst*, 12(7):706–715.e4, 2021. ISSN 2405-4712. doi: 10.1016/j.cels.2021.05.006.

Bernard J. Baars. A cognitive theory of consciousness. 1988.

T. Biancalani, G. Scalia, L. Buffoni, R. Avasthi, Z. Lu, A. Sanger, N. Tokcan, C. R. Vanderburg, Å Segerstolpe, M. Zhang, I. Avraham-Davidi, S. Vickovic, M. Nitzan, S. Ma, A. Subramanian, M. Lipinski, J. Buenrostro, N. B. Brown, D. Fanelli, X. Zhuang, E. Z. Macosko, and A. Regev. Deep learning and alignment of spatially resolved single-cell transcriptomes with tangram. *Nat Methods*, 18(11):1352–1362, 2021. ISSN 1548-7091 (Print) 1548-7091. doi: 10.1038/s41592-021-01264-7.

Charlotte Bunne, Yusuf Roohani, Yanay Rosen, and et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024. ISSN 0092-8674. doi: 10.1016/j.cell.2024.11.015. URL https://doi.org/10.1016/j.cell.2024.11.015.

D. M. Cable, E. Murray, L. S. Zou, A. Goeva, E. Z. Macosko, F. Chen, and R. A. Irizarry. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat Biotechnol*, 40(4):517–526, 2022. ISSN 1087-0156 (Print) 1087-0156. doi: 10.1038/s41587-021-00830-w.

H. Chen, E. Murray, A. Sinha, A. Laumas, J. Li, D. Lesman, X. Nie, J. Hotaling, J. Guo, B. R. Cairns, E. Z. Macosko, C. Y. Cheng, and F. Chen. Dissecting mammalian spermatogenesis using spatial transcriptomics. *Cell Rep*, 37(5):109915, 2021. doi: 10.1016/j.celrep.2021.109915.

J. Chen, H. Xu, W. Tao, Z. Chen, Y. Zhao, and J. J. Han. Transformer for one stop interpretable cell type annotation. *Nat Commun*, 14(1):223, 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-35923-4.

Y. Chen, Y. Liang, X. Luo, and Q. Hu. Oxidative resistance of leukemic stem cells and oxidative damage to hematopoietic stem cells under pro-oxidative therapy. *Cell Death Dis*, 11(4):291, 2020. doi: 10.1038/s41419-020-2488-y.

H. Cui, C. Wang, H. Maan, K. Pang, F. Luo, N. Duan, and B. Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nat Methods*, 2024. ISSN 1548-7091. doi: 10.1038/s41592-024-02201-0.

E. Dann, A. M. Cujba, A. J. Oliver, K. B. Meyer, S. A. Teichmann, and J. C. Marioni. Precise identification of cell states altered in disease using healthy single-cell references. *Nat Genet*, 55 (11):1998–2008, 2023. ISSN 1061-4036 (Print) 1061-4036. doi: 10.1038/s41588-023-01523-7.

C. De Donno, S. Hediyeh-Zadeh, A. A. Moinfar, M. Wagenstetter, L. Zappia, M. Lotfollahi, and F. J. Theis. Population-level integration of single-cell datasets enables multi-scale analysis across samples. *Nat Methods*, 20(11):1683–1692, 2023. ISSN 1548-7091 (Print) 1548-7091. doi: 10.1038/s41592-023-02035-2.

Alan DenAdel, Madeline Hughes, Akshaya Thoutam, Anay Gupta, Andrew W. Navia, Nicolo Fusi, Srivatsan Raghavan, Peter S. Winter, Ava P. Amini, and Lorin Crawford. Evaluating the role of pre-training dataset size and diversity on single-cell foundation model performance. *bioRxiv*, pp. 2024.12.13.628448, 2024. doi: 10.1101/2024.12.13.628448. URL http://biorxiv.org/content/early/2024/12/17/2024.12.13.628448.abstract.

C. Domínguez Conde, C. Xu, L. B. Jarvis, D. B. Rainbow, and et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594):eabl5197, 2022. ISSN 0036-8075 (Print) 0036-8075. doi: 10.1126/science.abl5197.

Feyza Duman Keles. On the computational complexity of self-attention. *arXiv*, 2209.04881, 2022.

M. Fasolino, G. W. Schwartz, A. R. Patil, A. Mongia, and et al. Single-cell multi-omics analysis of human pancreatic islets reveals novel cellular states in type 1 diabetes. *Nat Metab*, 4(2):284–299, 2022. ISSN 2522-5812. doi: 10.1038/s42255-022-00531-x.

Anirudh Goyal, Aniket Didolkar, Alex Lamb, Kartikeya Badola, Nan Rosemary Ke, Nasim Rahaman, Jonathan Binas, Charles Blundell, Michael C. Mozer, and Yoshua Coordination among neural modules through a shared global workspace. *ICLR*, 2021.

J. M. Granja, S. Klemm, L. M. McGinnis, A. S. Kathiria, A. Mezger, M. R. Corces, B. Parks, E. Gars, M. Liedtke, G. X. Y. Zheng, H. Y. Chang, R. Majeti, and W. J. Greenleaf. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat Biotechnol*, 37 (12):1458–1465, 2019. ISSN 1087-0156 (Print) 1087-0156. doi: 10.1038/s41587-019-0332-7.

Y. Hao, S. Hao, E. Andersen-Nissen, 3rd Mauck, W. M., and et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, 2021. ISSN 0092-8674 (Print) 0092-8674. doi: 10.1016/j.cell.2021.04.048.

S. He, R. Bhatt, C. Brown, E. A. Brown, D. L. Buhr, and et al. High-plex imaging of rna and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nat Biotechnol*, 40 (12):1794–1806, 2022. ISSN 1087-0156. doi: 10.1038/s41587-022-01483-z.

K. Hrovatin, A. Bastidas-Ponce, M. Bakhti, L. Zappia, M. Büttner, C. Salinno, M. Sterr, A. Böttcher, A. Migliorini, H. Lickert, and F. J. Theis. Delineating mouse -cell identity during lifetime and in diabetes with a single cell atlas. *Nat Metab*, 5(9):1615–1637, 2023. ISSN 2522-5812. doi: 10.1038/s42255-023-00876-x.

N. L. Jorstad, J. H. T. Song, D. Exposito-Alonso, H. Suresh, N. Castro-Pacheco, and et al. Comparative transcriptomics reveals human-specific cortical features. *Science*, 382(6667):eade9516, 2023. ISSN 0036-8075 (Print) 0036-8075. doi: 10.1126/science.ade9516.

T. Kumar, K. Nee, R. Wei, S. He, Q. H. Nguyen, and et al. A spatially resolved single-cell genomic atlas of the adult human breast. *Nature*, 620(7972):181–191, 2023. ISSN 0028-0836. doi: 10.1038/s41586-023-06252-9.

Yun Li, Zheng Huang, Lubin Xu, Yanling Fan, and et al. Uda-seq: universal droplet microfluidics-based combinatorial indexing for massive-scale multimodal single-cell sequencing. *Nature Methods*, 2025. ISSN 1548-7105. doi: 10.1038/s41592-024-02586-y. URL https://doi.org/10.1038/s41592-024-02586-y.

R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef. Deep generative modeling for single-cell transcriptomics. *Nat Methods*, 15(12):1053–1058, 2018. ISSN 1548-7091 (Print) 1548-7091. doi: 10.1038/s41592-018-0229-2.

M. Lotfollahi, M. Naghipourfar, M. D. Luecken, M. Khajavi, M. Büttner, M. Wagenstetter, Ž Avsec, A. Gayoso, N. Yosef, M. Interlandi, S. Rybakov, A. V. Misharin, and F. J. Theis. Mapping single-cell data to reference atlases by transfer learning. *Nat Biotechnol*, 40(1):121–130, 2022. ISSN 1087-0156 (Print) 1087-0156. doi: 10.1038/s41587-021-01001-7.

M. Lotfollahi, Hao Yuhan, F. J. Theis, and R. Satija. The future of rapid and automated single-cell data analysis using reference mapping. *Cell*, 187(10):2343–2358, 2024. ISSN 0092-8674. doi: 10.1016/j.cell.2024.03.009.

M. D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché, and F. J. Theis. Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods*, 19(1):41–50, 2022. ISSN 1548-7091 (Print) 1548-7091. doi: 10.1038/s41592-021-01336-8.

Q. Ma, Y. Jiang, H. Cheng, and D. Xu. Harnessing the deep learning power of foundation models in single-cell omics. *Nat Rev Mol Cell Biol*, 2024. ISSN 1471-0072. doi: 10.1038/s41580-024-00756-6.

S. Mages, N. Moriel, I. Avraham-Davidi, E. Murray, J. Watter, F. Chen, O. Rozenblatt-Rosen, J. Klughammer, A. Regev, and M. Nitzan. Tacco unifies annotation transfer and decomposition of cell identities for single-cell and spatial omics. *Nat Biotechnol*, 41(10):1465–1473, 2023. ISSN 1087-0156 (Print) 1087-0156. doi: 10.1038/s41587-023-01657-3.

I. Gabitto Mariano, J. Travaglini Kyle, M. Rachleff Victoria, and et al. Integrated multimodal cell atlas of alzheimer's disease. *bioRxiv*, pp. 2023.05.08.539485, 2023. doi: 10.1101/2023.05.08.539485. URL http://biorxiv.org/content/early/2023/06/16/2023.05.08.539485.abstract.

J. R. Moffitt, D. Bambah-Mukku, S. W. Eichhorn, E. Vaughn, K. Shekhar, J. D. Perez, N. D. Rubinstein, J. Hao, A. Regev, C. Dulac, and X. Zhuang. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*, 362(6416), 2018. ISSN 0036-8075 (Print) 0036-8075. doi: 10.1126/science.aau5324.

R. S. Negrin. Graft-versus-host disease versus graft-versus-leukemia. *Hematology Am Soc Hematol Educ Program*, 2015:225–30, 2015. ISSN 1520-4383. doi: 10.1182/asheducation-2015.1.225.

G. Novakovsky, N. Dexter, M. W. Libbrecht, W. W. Wasserman, and S. Mostafavi. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nat Rev Genet*, 24(2):125–137, 2023. ISSN 1471-0056. doi: 10.1038/s41576-022-00532-2.

Boiarsky Rebecca, Singh Nalini, Buendia Alejandro, Getz Gad, and Sontag David. A deep dive into single-cell rna sequencing foundation models. *bioRxiv*, pp. 2023.10.19.563100, 2023. doi: 10.1101/2023.10.19.563100. URL http://biorxiv.org/content/early/2023/10/23/2023.10.19.563100.abstract.

S. G. Rodriques, R. R. Stickels, A. Goeva, C. A. Martin, E. Murray, C. R. Vanderburg, J. Welch, L. M. Chen, F. Chen, and E. Z. Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, 2019. ISSN 0036-8075 (Print) 0036-8075. doi: 10.1126/science.aaw1219.

Yanay Rosen, Yusuf Roohani, Ayush Agarwal, Leon Samotorčan, Consortium Tabula Sapiens, Stephen R. Quake, and Jure Leskovec. Universal cell embeddings: A foundation model for cell biology. *bioRxiv*, pp. 2023.11.28.568918, 2023. doi: 10.1101/2023.11.28.568918. URL http://biorxiv.org/content/early/2023/11/29/2023.11.28.568918.abstract.

A. J. C. Russell, J. A. Weir, N. M. Nadaf, M. Shabet, V. Kumar, S. Kambhampati, R. Raichur, G. J. Marrero, S. Liu, K. S. Balderrama, C. R. Vanderburg, V. Shanmugam, L. Tian, J. B. Iorgulescu, C. H. Yoon, C. J. Wu, E. Z. Macosko, and F. Chen. Slide-tags enables single-nucleus barcoding for multimodal spatial genomics. *Nature*, 625(7993):101–109, 2024. ISSN 0028-0836 (Print) 0028-0836. doi: 10.1038/s41586-023-06837-4.

S. Salcher, G. Sturm, L. Horvath, G. Untergasser, C. Kuempers, G. Fotakis, E. Panizzolo, A. Martowicz, M. Trebo, G. Pall, G. Gamerith, M. Sykora, F. Augustin, K. Schmitz, F. Finotello, D. Rieder, S. Perner, S. Sopper, D. Wolf, A. Pircher, and Z. Trajanoski. High-resolution single-cell atlas reveals diversity and plasticity of tissue-resident neutrophils in non-small cell lung cancer. *Cancer Cell*, 40(12):1503–1520.e8, 2022. ISSN 1535-6108 (Print) 1535-6108. doi: 10.1016/j.ccell.2022.10.008.

L. Steuernagel, B. Y. H. Lam, P. Klemm, G. K. C. Dowsett, C. A. Bauder, J. A. Tadross, T. S. Hitschfeld, A. Del Rio Martin, W. Chen, A. J. de Solis, H. Fenselau, P. Davidsen, I. Cimino, S. N. Kohnke, D. Rimmington, A. P. Coll, A. Beyer, G. S. H. Yeo, and J. C. Brüning. Hypomap-a unified single-cell gene expression atlas of the murine hypothalamus. *Nat Metab*, 4(10):1402–1419, 2022. ISSN 2522-5812. doi: 10.1038/s42255-022-00657-y.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06139-9. URL https://doi.org/10.1038/s41586-023-06139-9.

Liu Tianyu, Li Kexing, Wang Yuge, Li Hongyu, and Zhao Hongyu. Evaluating the utilities of large language models in single-cell data analysis. *bioRxiv*, pp. 2023.09.08.555192, 2023. doi: 10.1101/2023.09.08.555192. URL http://biorxiv.org/content/early/2023/09/08/2023.09.08.555192.abstract.

P. van Galen, V. Hovestadt, M. H. Wadsworth Ii, T. K. Hughes, G. K. Griffin, S. Battaglia, J. A. Verga, J. Stephansky, T. J. Pastika, J. Lombardi Story, G. S. Pinkus, O. Pozdnyakova, I. Galinsky, R. M. Stone, T. A. Graubert, A. K. Shalek, J. C. Aster, A. A. Lane, and B. E. Bernstein. Single-cell rna-seq reveals aml hierarchies relevant to disease progression and immunity. *Cell*, 176(6): 1265–1281.e24, 2019. ISSN 0092-8674 (Print) 0092-8674. doi: 10.1016/j.cell.2019.01.031.

Kirill Vishniakov, Karthik Viswanathan, Aleksandr Medvedev, Praveen K. Kanithi, Marco A. F. Pimentel, Ronnie Rajan, and Shadab Khan. Genomic foundationless models: Pretraining does not promise performance. *bioRxiv*, pp. 2024.12.18.628606, 2024. doi: 10.1101/2024.12.18.628606. URL http://biorxiv.org/content/early/2024/12/20/2024.12.18.628606.abstract.

A METHODS

A.1 THE METHODOLOGY OF CELLMEMORY

Embedding: Position embedding is used to characterize features such as genes, while gene expression is processed in a bag-of-words manner and characterized by token embedding (rounding up or scaling to bin ranges, e.g., 50).

Construct specialist matrix. For each cell, $n \in N$ of the input matrix $\boldsymbol{X}$, an extra CLS token is appended to the front of genes. This extended representation is further converted into a K-dimensional embedding space, resulting in the specialist matrix $\boldsymbol{R} \in \mathbb{R}^{N\times(1+M)\times K}$, where K is a hyperparameter identifying the number of dimensions essential for processing complex data relationships, and M represents the number of specialists (genes).

Compete to write. The specialist modules $\boldsymbol{R}$ compete to write the information to the shared global workspace, resulting in the memory matrix $\boldsymbol{D} \in \mathbb{R}^{N\times H\times K}$, with H distinct memory slots serving as a hyperparameter. Typically, H is much smaller than the original input dimension M, reflecting the bottleneck of the global workspace. The writing procedure involves a cross-attention mechanism between the specialist matrix $\boldsymbol{R}$ and the memory matrix $\boldsymbol{D}$, detailed as follows.

$\tilde{\boldsymbol{W}}_q$, $\tilde{\boldsymbol{W}}_k$, and $\tilde{\boldsymbol{W}}_v$ represent the respective weight matrices for the Query, Key, and Value when calculating the cross-attention matrix.

$$\begin{aligned} \text{Query:} \quad & \tilde{\boldsymbol{Q}} = \boldsymbol{D}\tilde{\boldsymbol{W}}_q \\ \text{Key:} \quad & \tilde{\boldsymbol{K}} = \boldsymbol{R}\tilde{\boldsymbol{W}}_k, \\ \text{Value:} \quad & \tilde{\boldsymbol{V}} = \boldsymbol{R}\tilde{\boldsymbol{W}}_v \end{aligned} \tag{1}$$

Then, the attention matrix $\tilde{\boldsymbol{A}}$ is calculated by multiplying the Query and Key matrices, scaled by the last dimension of attention weights d, and applying a softmax function for normalization. We select the top k largest attention scores along the $1 + M$ dimension, where specialized modules compete to write relevant information to the global workspace. It's worth noting that, in our experiments, we observed a slight performance improvement when using the top k mechanism, but it impacted the interpretability of the model results. Therefore, in this article application, we did not employ the top k.

$$\tilde{\boldsymbol{A}} = \text{Top}_k(\text{softmatx}(\frac{\tilde{\boldsymbol{Q}}\tilde{\boldsymbol{K}}^T}{\sqrt{d}})), \tag{2}$$

Finally, the attention matrix$\tilde{\boldsymbol{A}}$ is multiplied by the Value matrix $\tilde{\boldsymbol{V}}$, resulting in the updated memory $\boldsymbol{D}$.

$$\boldsymbol{D} = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{V}}, \tag{3}$$

Broadcast to specialists. The memory matrix $\boldsymbol{D}$ is broadcasted back to the specialists, leading to the creation of an updated specialist matrix $\boldsymbol{R}$. This broadcasting action also utilizes a cross-attention mechanism, which is demonstrated as follows. (1) Similar to the writing mechanism, we

formulate the Query, Key, and Value matrices when computing the cross-attention for broadcasting. Specifically, the weight matrices associated with the Query, Key, and Value are represented by $\hat{\boldsymbol{W}}_q$, $\hat{\boldsymbol{W}}_k$, and $\hat{\boldsymbol{W}}_v$ respectively.

$$\begin{aligned} \text{Query:} \quad & \hat{\boldsymbol{Q}} = \boldsymbol{R}\hat{\boldsymbol{W}}_q \\ \text{Key:} \quad & \hat{\boldsymbol{K}} = \boldsymbol{D}\hat{\boldsymbol{W}}_k, \\ \text{Value:} \quad & \hat{\boldsymbol{V}} = \boldsymbol{D}\hat{\boldsymbol{W}}_v \end{aligned} \tag{4}$$

Then, we calculate the attention matrix $\hat{\boldsymbol{A}}$ by applying softmax over the slots.

$$\hat{\boldsymbol{A}} = \text{softmatx}(\frac{\hat{\boldsymbol{Q}}\hat{\boldsymbol{K}}^T}{\sqrt{d}}), \tag{5}$$

Eventually, we construct the updated specialist matrix $\boldsymbol{R}$.

$$\boldsymbol{R} = \hat{\boldsymbol{A}}\hat{\boldsymbol{V}}, \tag{6}$$

Model training and prediction. In the training process of CellMemory, the CLS token matrix $\boldsymbol{S} \in \mathbb{R}^{N \times K}$ is extracted from the first dimension of the revised specialist matrix $\boldsymbol{R}$ at index 0: $S_(n, k) = R_(n, 0, k)$. Through the iterative process of writing and broadcasting, the CLS token acquires substantial intracellular information. The CLS token $\boldsymbol{S}$ is then projected through a fully connected layer f and processed to the probability for each class $c \in \boldsymbol{C}$ using softmax function. The predictions are compared to the true labels, and the cross-entropy loss $\mathcal{L}_{CE}$ is calculated as follows.

$$\begin{aligned} \hat{\boldsymbol{y}} &= \text{softmax}(f(\boldsymbol{S})) \\ \mathcal{L}_{CE} &= -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_{n,c} \log \hat{y}_{n,c} \end{aligned} \tag{7}$$

CellMemory is guided to focus on acquiring information related to the identity representation of cells. During the prediction, the softmax function assigns each cell's prediction results to the probability of each category, with the category having the highest probability (or confidence score) being considered as the cell's identity.

A.2 Hyperparameter settings

The gene expression matrix, after undergoing embedding, is input into a 4-layered transformer along with the CLS token for classification. We set the embedding dimension (*K*) to 256 and the feed-forward network dimension to 512. The Query and Key sizes are set to 32, and the Value size is set to 64. We employ 4 heads during reading from and writing into the global workspace, which comprises 8 memory slots (*H*). We train the model using an early stopping strategy, stopping training when the validation set's loss has not decreased for 5 epochs. We use Adam optimizer with a learning rate of 0.0003 and anneal the learning rate using cosine annealing.

A.3 Performing downstream tasks

CLS token: The CLS token $\boldsymbol{S} \in \mathbb{R}^{N \times K}$ effectively communicates with gene tokens inside the cell through attention mechanisms and the bottleneck architecture in a supervised manner, obtaining a complete cellular representation. It is employed for single-cell and single-cell spatial data reference mapping.

Attention / Memory score matrix: The interpretability of CellMemory originates from the cross-attention mechanism's explicit demonstration of the importance of input features. The gene expression matrix undergoes a writing and broadcasting process with the memory, leading to the generation of two memory score matrices ($\tilde{\boldsymbol{A}} \in \mathbb{R}^{N \times H \times (1+M)}$ and ($\hat{\boldsymbol{A}} \in \mathbb{R}^{N \times (1+M) \times H}$). Matrix multiplication is performed on the two memory matrices derived from the writing and broadcasting steps, yielding a resultant matrix $\boldsymbol{Z} = \hat{\boldsymbol{A}}\tilde{\boldsymbol{A}} \in \mathbb{R}^{N \times (1+M) \times (1+M)}$. From matrix $\boldsymbol{Z}$, we extract matrix $\boldsymbol{Z} \in \mathbb{R}^{N \times M}$, where $\boldsymbol{Z}'_{n,m} = \boldsymbol{Z}_{n,1,m+1}$, capturing the attention of each CLS token concerning each

of the M genes across N cells. To generate cell-type-specific TAGs (from attention score matrix), the averaged attention score for specific groups of cells is computed based on the **global attention score matrix $\boldsymbol{Z}$**. Then, the scores are sorted to obtain the top attention genes by memory score for each group. To generate memory slot specific TAGs, the **memory score matrix $\hat{\boldsymbol{A}}$** produced during broadcasting is used. The mean is calculated across selected cells, then, based on averaged attention scores, the features are sorted to obtain TAGs for each memory slot.

i) Interpretation Level 1 (gene level): To generate cell-type-specific TAGs, we compute the average attention score for specific groups of cells based on the global attention score matrix Z. Then, the scores are sorted to obtain the top attention genes for each group. In this manner, we obtain the feature attribution score for each cell. Typically, we observe that, within a given cell type, the attention scores of the top 20 genes exhibit clear consistency with gene expression levels. When systematically analyzing the functional pathways associated with these top attention genes, we recommend considering a larger gene set, such as the top 50 or 100 genes.

ii) Interpretation Level 2 (gene program level): To generate memory slot specific TAGs, we utilize the memory matrix A produced during broadcasting. The mean is calculated across selected cells, then, based on averaged memory scores, the features are sorted to obtain TAGs for each slot. At this level, it becomes possible to determine, for each cell, which features each slot within the model focuses on the most. When analyzing the functional pathway enrichment of genes associated with a memory slot, we suggest using a gene set comprising the top 50 genes.

A.4 THEORETICAL ANALYSIS

CellMemory is more computationally efficient than the conventional self-attention mechanisms, especially in single-cell omics data analysis. Specifically, the traditional self-attention approach used in BERT is characterized by quadratic complexity as a function of sequence length. This inherent complexity becomes a substantial impediment as sequences extend, particularly when dealing with voluminous single-cell omics datasets. Using traditional self-attention for such extensive data requires substantial computational resources and elongated processing time, making it unsuitable for real-time applications. In contrast, CellMemory harnesses a cross-attention mechanism to implement the shared global workspace, significantly reducing time and space complexity.

Time complexity. Our proposed method, CellMemory, outperforms traditional self-attention mechanisms in terms of computational time complexity. Consider the "Compete to Write" procedure. For each cell $n \in N$, we have Query $\tilde{\boldsymbol{Q}}_n \in \mathbb{R}^{H \times d}$ and Key $\tilde{\boldsymbol{K}}_n \in \mathbb{R}^{(1+M) \times d}$, and the total time complexity to calculate cross-attention matrix $\tilde{\boldsymbol{A}}_n$ is $\mathcal{O}(MHd)$ in our proposed method. Similarly, due to the consistent use of cross-attention, the time complexity to calculate $\tilde{\boldsymbol{A}}_n$ is also $\mathcal{O}(MHd)$ in the "Broadcast to Specialists" procedure. In comparison, the traditional self-attention mechanism exhibits a total time complexity of $\mathcal{O}(M^2 d)$. Specifically, H, the hyperparameter representing memory slots, is typically chosen to be much smaller than the input dimension *M* to reflect the bottleneck of the shared global workspace.

Space complexity. Furthermore, CellMemory also surpasses the traditional self-attention mechanisms regarding space complexity. For each cell $n \in N$, the "Compete to Write" procedure yields a cross-attention matrix $\tilde{\boldsymbol{A}}_n$ with dimensions $H \times (1 + M)$, while the "Broadcast to Specialists" procedure results in $\tilde{\boldsymbol{A}}_n$ with dimensions $(1 + M) \times H$. Consequently, the space complexity of our method is $\mathcal{O}(MH)$. In contrast, the traditional self-attention mechanisms operate with a quadratic space complexity of $\mathcal{O}(M^2)$. Given $H \ll M$, it is evident that CellMemory offers a more memory-efficient alternative, saving valuable computational resources compared to the conventional self-attention techniques.

A.5 RELATION WITH OTHER TRANSFORMER VARIANTS

Transformers, with their attention-based architecture, have become a cornerstone in deep learning. However, traditional transformers face challenges like quadratic time and memory complexity and a large parameter count, impeding scalability. To address these issues, numerous transformer variants have emerged in recent years. Our method aligns with variants incorporating a 'Neural Memory' and implementing down-sampling within the attention mechanism. This approach is directly inspired by the transformer shared global workspace and bears similarities to the Perceiver model.

Both strategies refine the pairwise self-attention mechanism by introducing global coordination and a unified memory. This integrated memory acts as a constrained communication channel, distinct from other transformer bottleneck techniques like discrete bottleneck and restricted cross-modal attention. In our model, bottleneck effects are achieved through a small, integrated memory, facilitating information integration and dimension reduction. These features are particularly advantageous for processing complex and noisy biological data, as in our study.

B DATASETS

B.1 SINGLE-CELL DATASET

hBreast. We collected a single cell atlas of the adult human breast from the CELLxGENE database. This dataset includes 244,285 cells (100,000 cells from normal samples by down-sampling; 144,285 cells from breast cancer samples), identifying 58 biological cell states. The data was subset to the 3,000 highly variable genes (HVGs) to benchmark. We established an annotation scenario using normal cells as the reference for annotating cells from breast cancer samples. The CLS embedding of the query set was evaluated in the integration benchmark. Batch information was denoted using the 'library_uuid'.

hLung. The Lung atlas is available in the CELLxGENE database, comprising both normal and LUSC samples It contains 305,319 cells (212,889 cells from normal samples and 92,430 cells from LUSC samples) and identifies 24 cell types. We restricted the reference dataset to data exclusively from the 10x platform to predict the cell type of the query set of LUSC. The query set incorporated data from 5 platforms. In the integration benchmark, the term 'dataset' was utilized as the batch label. A total of 3,000 HVGs were used for benchmarking.

hPancreas. The human pancreas dataset includes 47,644 cells (26,197 cells from control samples; 21,447 cells from T1D patients, and 13 cell types). We removed the hybrid cell type from the dataset, and cells from control donors were used as the reference while cells from T1D patients were analyzed as the query set. A total of 3,000 HVGs were used for benchmarking.

Immune. The training set was composed of 141,524 immune cells derived from 8 tissues, including the lung and liver. The test set consisted of 188,238 immune cells from 9 tissues, such as the spleen and thymus. A total of 4,000 HVGs were used for benchmarking.

mPancreas. The mPancreas contains scRNA-seq data from 239,679 mouse pancreatic islet cells (with the reference down-sampled to 100,000). The normal cells within the dataset served as a reference to train the model, and the T1/2D samples were utilized as the query set (139,679 cells). This dataset comprises 20 cell types. The term 'batch_integration' was utilized as the batch label. A total of 3,000 HVGs were used for benchmarking.

mHypoMap. The mouse hypothalamic atlas from HypoMap contains 66 cell populations derived from the transcriptional profiling of 161,903 cells. We used the data from the Drop-seq platform as the query set (61,903 cells) and the data from the 10x platform as the train set (downsampled to 100,000 cells). The term 'Sample_ID' was utilized as the batch label. A total of 4,000 HVGs were used for benchmarking.

crossSpecies (cortex). In the cross-species benchmark, we utilized human cerebral cortex data, comprising 200,000 cells, as the training set. The query set was constituted by a dataset that includes one human and four non-human cerebral cortex datasets (human: 156,285 cells, Gorilla: 139,945 cells; Chimpanzee: 112,929 cells; Macaque: 89,136 cells; Marmoset: 75,861 cells). This part ensured the retention of a set of homologous genes across all species. The performance of the integration was evaluated using chimpanzee cortex data. A total of 3,000 HVGs were used for benchmarking.

MPAL lineage inference. Bone marrow and peripheral blood mononuclear cells from healthy samples were used, totaling 35,582 cells (MBBCs=12,602; CD34+- enriched BMMCs=8,176; PBMCs=14,804). A model was constructed using 1,000 highly variable genes retained from the reference. MPAL-a (3,435 cells) corresponds to B-myeloid MPAL (MPAL4) in the original study, and MPAL-b (2,947 cells) corresponds to T-myeloid MPAL (MPAL5). Cells labeled as healthy-like in the original study were excluded.

B.2 SINGLE-CELL SPATIAL DATASET

mHypo. The mouse hypothalamic spatial dataset measured by MERFISH, comprising three samples with naïve behavior, totaling 180,754 cells. The single-cell data from mHypoMap was used as the reference. Aligning genes (154) were captured by MERFISH resulting in 8 categories.

hNSCLC. The scRNA-seq dataset of hLung served as the reference to predict NSCLS spatial dataset. The NSCLC spatial dataset includes Lung-6, Lung-9, and Lung-13, consisting of 256,581 cells from 11 cell types. These samples were generated from a Formalin-Fixed Paraffin-Embedded (FFPE) sample by the CosMx SMI platform. A shared gene set of 952 genes from the reference dataset was selected to align with the query dataset.

mSpermato. The spatial dataset of mouse spermatogenesis (measured by Slide-seq) was acquired from three wild-type (WT) mice and three leptin-deficient diabetic (ob/ob) mice, including 207,335 cells and 9 cell types. The WT dataset served as the reference, and the ob/ob dataset served as the query set.

Slide-tags human cortex. The spatial transcriptome data generated by Slide-tags sequencing consists of 4,065 cells belonging to 7 cell types. Cells from normal samples in the brain cell atlas of Alzheimer's disease were utilized as a reference for integrating the Slide-tags spatial data.