BIOLOGICALLY INTERPRETABLE LATENT REPRESENTATIONS IN SINGLE-CELL MULTIMODALITIES

by

Sayali Anil Alatkar

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Computer Science)

at the

UNIVERSITY OF WISCONSIN-MADISON

2025

Date of Final Oral Exam: 08/04/2025

The dissertation is approved by the following members of the Final Oral Committee:

Daifeng Wang (advisor), Associate Professor, Biostatistics and Medical Informatics
Frederic Sala, Assistant Professor, Computer Science
Yudong Chen, Associate Professor, Computer Science
Andre Sousa, Assistant Professor, Neuroscience
Panos Roussos, Professor, Psychiatry and Genetics and Genomic Sciences, (Icahn School of Medicine at Mount Sinai)

Biologically interpretable latent representations in single-cell multimodalities

Sayali Anil Alatkar

Abstract

High-dimensional, multimodal single-cell data are reshaping our understanding of cellular systems, offering unprecedented insights into cellular diversity, molecular dynamics, and disease mechanisms. Yet their full potential remains constrained by critical gaps: technical artifacts (e.g., missing modalities), the static snapshot nature of sequencing obscuring dynamic trajectories, and the absence of personalized frameworks for individual-level analysis. This thesis bridges these gaps by developing biologically interpretable computational frameworks for multimodal single-cell genomics, enabling robust integration, dynamic reconstruction, and personalized disease profiling.

First, we introduce Cross-Modality Optimal Transport (CMOT) to impute missing modalities by integrating nonlinear manifold alignment, entropy-regularized optimal transport, and k-nearest neighbors. Validated across neurodevelopment, cancer, and immunology contexts, CMOT outperforms state-of-the-art tools in preserving biological variation, retaining cell-type markers in brain development, and recovering treatment-response genes in oncology. Its flexibility to incorporate prior knowledge and scalability to small datasets broadens applicability.

Second, to resolve dynamic processes from static snapshots, we develop ARTEMIS, a generative model coupling variational autoencoders with unbalanced Schrödinger Bridges. By embedding cells into a latent space governed by stochastic differential equations, ARTEMIS reconstructs continuous trajectories and population shifts across time-series data. In pancreatic development, zebrafish embryogenesis, and the epithelial-mesenchymal transition, it achieves superior trajectory accuracy, identifies driver genes, and scales optimally to high-dimensional genomics.

Finally, addressing personalized disease mechanisms, we develop iBrainMap, a knowledge-guided graph attention network (GAT) framework. By analyzing millions of nuclei from Alzheimer's disease brains, it integrates individual-specific regulatory networks and cell-cell interactions, prioritizing genes linked to AD progression via diffusion kernels. This approach has uncovered disease population subgroups, trajectories, and molecular drivers associated with neurodegenerative disorders.

To my family and friends, who supported me throughout this journey.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my advisor, Daifeng Wang. Thank you for your steady mentorship, thoughtful guidance, and for giving me the space to explore my own ideas. You have taught me not only how to be a better scientist, but also how to think more clearly and communicate more thoughtfully. I've grown immensely under your guidance, and I'll carry many of your lessons with me for years to come.

Ting Jin, Pramod Chandrashekhar, and Noah Kalafut, I spent a significant part of my PhD working closely with you, and I could not have asked for better collaborators. Thank you for your creativity, dedication, and for always being generous with your time. I learned so much from each of you, from the science, yes, but also from the way you work, think, and support others.

To the rest of my labmates and collaborators, Chenfeng, Kalpana, Pubudu, Jei, Jerome, and Chirag, thank you for being part of this journey. I have appreciated your energy, curiosity, and companionship, as well as all our scientific discussions.

I would also like to thank my committee members, Fred Sala and Yudong Chen, for their support and guidance throughout this process. I am also deeply grateful to Panos Roussos for his guidance and continued support as a senior collaborator on multiple projects; your scientific insight, thoughtful feedback, and generosity with your time have been invaluable. I am also grateful to Andre Sousa for the opportunity to collaborate on a project together, your feedback and time have been very much appreciated.

I am especially thankful to Saniya Khullar for being a constant presence throughout my PhD, for the many honest conversations, shared frustrations, and thoughtful reflections on science and life. You helped me stay grounded through the highs and lows, and I have cherished our friendship along the way.

Outside the lab, I am thankful to Shivani for always being there, for listening without judgment, for thoughtful conversations, and for reminding me to zoom out and breathe when I needed it most.

To my family, thank you for everything. My parents, Dr. Anita and Dr. Anil, your love, strength, and confidence in me have been the foundation of all my pursuits. Tushar, thank you for always being the calm voice of reason. Omik, thank you for your love, patience, and unwavering belief in me. This journey would have looked very different without you. To my amazing in-laws, Asha and Gokul Mahajan, and to Smitesh and Harsha, thank you for your constant support and words of

encouragement.

To everyone who walked beside me in small and big ways, thank you. You made this possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The study of single-cell biology has transformed our understanding of cellular identity, development, and function. Traditional bulk profiling methods average over heterogeneous populations, obscuring the diversity of cell types and states. In contrast, single-cell technologies enable high-resolution profiling of gene expression, chromatin accessibility, and other molecular modalities at the level of individual cells. These advances have opened new opportunities to dissect developmental processes, tissue composition, and disease mechanisms at an unprecedented scale and resolution.

Recent large-scale single-cell studies have produced multimodal datasets profiling hundreds of thousands to millions of cells across diverse tissues, timepoints, and individuals. For example, single-nucleus RNA-seq (snRNA-seq) captures expression of over 20,000 genes per nucleus, while assays such as scATAC-seq or CITE-seq extend profiling to chromatin accessibility and surface protein abundance, respectively. These datasets have enabled the construction of comprehensive cellular atlases across developmental and disease contexts. However, they also present significant computational challenges due to their high dimensionality, sparsity, and the destructive nature of single-cell profiling, which limits joint measurements across modalities or time.

Three core analytical challenges limit the full potential of these data. First, the destructive nature of current protocols results in missing modalities, i.e., each cell is typically measured for only one molecular layer, making integrated analysis difficult. Second, most datasets provide snapshot measurements of dynamic processes, limiting our ability to reconstruct the temporal evolution of cell states. Third, as studies increasingly scale to human populations, modeling inter-individual heterogeneity becomes essential: capturing how molecular programs vary across individuals is critical for understanding disease mechanisms and developing personalized approaches.

To address the first challenge, we developed CMOT (Cross-Modality Optimal Transport), a computational framework for aligning and integrating single-cell datasets across different molecular modalities [4]. CMOT leverages optimal transport to align distributions of unpaired cells, learning a shared latent space that preserves biological structure while enabling accurate imputation. It accommodates modality-specific sparsity and partial overlap, offering a flexible and scalable solution for multimodal

data integration across diverse experimental settings.

While CMOT helps bridge disconnected molecular views of the same system, it still provides a static snapshot of cellular organization. Yet many biological processes, such as development, differentiation, or disease progression, are inherently dynamic, unfolding over time as cells transition between states. To capture this temporal dimension, we introduced ARTEMIS, a probabilistic trajectory inference framework based on unbalanced diffusion Schrödinger bridges [5]. ARTEMIS models time-evolving cell state distributions by estimating both drift and mass-change terms from temporally indexed data. It combines a variational autoencoder with learned stochastic dynamics to capture both uncertainty and directionality in cellular transitions. This enables biologically meaningful reconstructions of differentiation paths and disease progression trajectories, overcoming limitations of deterministic or static models.

Finally, while earlier parts of this thesis focus on understanding how cells function, change over time, and integrate information across molecular layers, it is equally important to widen our perspective from the cellular level to the level of the human. In human biology, cellular programs are shaped by the broader context of each individual, including genetics, environment, and disease state. This inter-individual variability is especially important in complex disorders like Alzheimer's, where similar clinical symptoms can arise from distinct molecular mechanisms. As single-cell studies increasingly scale to human populations, there is a critical need for personalized frameworks that can capture how gene regulation and cellular interactions vary across individuals. To address this, we developed iBrainMap, a framework for personalized functional genomics in Alzheimer's disease [24]. Using a knowledge-guided graph neural network, iBrainMap learns latent embeddings that preserve each donor's unique molecular context. These individualized representations enable improved phenotype classification, subtype discovery, and inference of disease trajectories at the person-specific level. By modeling how cells and their interactions vary across individuals, iBrainMap offers a scalable approach for studying human heterogeneity in single-cell data and contributes to the development of precision neuroscience.

Together, the methods developed in this thesis, CMOT, ARTEMIS, and iBrainMap, tackle core challenges in multimodal integration, dynamic modeling, and personalized inference in single-cell genomics. These approaches contribute to a deeper understanding of how cellular programs emerge, evolve, and vary across individuals. Their utility spans a broad range of biological applications, including embryonic development, neurodegeneration, and immune response. A further discussion of the broader implications of this work, along with future directions, is provided at the end of this thesis.

Publications

The work presented in this thesis is based on the following peer-reviewed publications and manuscripts under revision. Each chapter reflects one of these contributions:

- Sayali Anil Alatkar, Daifeng Wang. CMOT: Cross-Modality Optimal Transport for multimodal inference. *Genome Biol.* 24, 163, 2023. [4]
- Sayali Anil Alatkar, Daifeng Wang. ARTEMIS integrates autoencoders and schrödinger bridges to predict continuous dynamics of gene expression, cell population and perturbation from time-series single-cell data. In press, *Bioinformatics*, ECCB, 2025. [5]
- Pramod Bharadwaj Chandrashekar*, Sayali Anil Alatkar*, Noah Cohen Kalafut*, Ting Jin*, Chirag Gupta, Ryan Burzak, Xiang Huang, Shuang Liu, Athan Z. Li, PsychAD Consortium, Kiran Girdhar, Georgios Voloudakis, Gabriel E. Hoffman, Jaroslav Bendl, John F. Fullard, Donghoon Lee, Panos Roussos, Daifeng Wang. Personalized Single-cell Transcriptomics Reveals Molecular Diversity in Alzheimer's Disease. *in revision*, 2025. [24]

Further Publications

The following publications of the author and collaborators are more broadly relevant to the topic of this thesis but have not been directly included:

- Chirag Gupta, Jielin Xu, Ting Jin, Saniya Khullar, Xiaoyu Liu, Sayali Alatkar, Feixiong Cheng, and Daifeng Wang. Single-cell network biology characterizes cell type gene regulation for drug repurposing and phenotype prediction in Alzheimer's disease. *PLoS Computational Biology* 18, no. 7, 2022: e1010287. [59]
- Pramod Bharadwaj Chandrashekar, Sayali Alatkar, Jiebiao Wang, Gabriel E. Hoffman, Chenfeng He, Ting Jin, Saniya Khullar, Jaro Bendl, John F. Fullard, Panos Roussos, Daifeng Wang. DeepGAMI: deep biologically guided auxiliary learning for multimodal integration and imputation to improve genotype–phenotype prediction. *Genome Medicine* 15, no. 1, 2023: 88. [23]
- Xiang Huang, Noah Cohen Kalafut, Sayali Alatkar, Athan Z. Li, Qiping Dong, Qiang Chang, and Daifeng Wang. NeuroTD: A Time-Frequency Based Multimodal Learning Approach to Analyze Time Delays in Neural Activities. *bioRxiv*, 2024: 2024-10. [76]

Collaborators

This thesis would not have been possible without the guidance of my advisor, Daifeng Wang. Many of the ideas presented here were shaped through our discussions and meetings. I also benefited greatly from collaborations with colleagues, and unless otherwise cited, the results presented are the work of the author and collaborators. In particular, Chapter 4 includes material from a research project with shared first authorship between the author, Pramod Chandrashekar, Noah Kalafut, and Ting Jin.

Chapter 2

Multimodal Integration and Imputation

2.1 Abstract

Multimodal measurements of single-cell sequencing technologies facilitate a comprehensive understanding of specific cellular and molecular mechanisms. However, simultaneous profiling of multiple modalities of single cells is challenging, and data integration remains elusive due to missing modalities and cell–cell correspondences. To address this, we developed a computational approach, Cross-Modality Optimal Transport (CMOT), which aligns cells within available multi-modal data (source) onto a common latent space and infers missing modalities for cells from another modality (target) of mapped source cells. CMOT outperforms existing methods in various applications from developing brain, cancers to immunology, and provides biological interpretations improving cell-type or cancer classifications.

2.2 Background

Single-cell sequencing technologies can measure different characteristics of single cells across multi-omics such as the genomics, transcriptomics, epigenomics, proteomics. Such high-resolution measurements have enabled exploring individual cells to reveal cellular and molecular mechanisms and study cell-to-cell functional variations. For example, scCAT-seq, sci-CAR, and 10xMultiome measure single-cell gene expression and chromatin accessibility [149], [17], and CITE-seq measures gene and protein expression of single cells [138], [54]. However, simultaneous profiling of such multi-omics and additional modalities continues to be a challenging task especially because of high sequencing costs, low recovery of individual cells, and sparse and noisy data [42]. Owing to these challenges, single-cell multimodal data generation may not always be feasible. This leads to the question of how we can use available multi-modalities to infer missing modalities.

Several prior works have tackled modality inference. Seurat [139, 66] infers the missing modality of a cell by weighting nearest neighboring cells with multimodalities available. MOFA+ [8] uses Bayesian factor analysis to identify a lower dimensional representation of the data to infer the missing modality. However, they only work with multimodal data that must come from the same cells (i.e., fully correspond-

ing). Alignment-based methods like non-linear manifold alignment [75] have shown to align multimodalities with partial cell-to-cell correspondence information but not been extended to cross-modality inference. Machine learning has also emerged to help modality inference. For instance, TotalVI [54] builds a variational autoencoder that infers missing protein profiles from gene expression using CITE-seq data. Polarbear [167] also uses autoencoders, however, trains on both single and multi-modal data to infer each modality. However, such autoencoder-based approaches are unsupervisedthat learn the latent embeddings that likely lack biological interpretability and lack a mechanism to introduce prior knowledge about underlying data distribution [128]. Moreover, training autoencoders typically requires considerable amounts of data and time with intensive hyperparameter tuning.

Optimal Transport (OT), an efficient approach uses prior knowledge about data distribution to find an optimal mapping between the distributions [120]. OT can also work on small datasets with limited parameters. Recently, OT has been applied to single-cell multiomics data for various applications [134], [39], [40], [19]. Schiebinger et al. [134] used OT to model the developmental trajectory of single-cell gene expression through unbalanced optimal transport. Single-cell integrative analysis frameworks like SCOT [39], SCOTv2 [40], and Pamona [19] further extended the original OT problem for multi-omics data alignment. Another work [78] used OT with an additional entropic regularization term to improve the unsupervised clustering of single-cell data to understand cell types and cellular states better. However, OT has not yet been applied for cross-modality inference. Thus, we propose that integrating OT with multimodal data alignment can work for cross-modality inference and address the above limitations of prior works.

Particularly, we developed CMOT (Cross-Modality Optimal Transport), a computational approach to infer missing modalities of single cells. CMOT first aligns the cells with multimodal data (source) if the cells do not have complete correspondence, and then applies OT to map the cells from single modality (target) to the source cells via shared modality. Finally, CMOT uses the k-Nearest-Neighbors (kNN) of source cells to infer missing modality for target cells. Moreover, CMOT does not need paired multi-modal data for alignment. We found that not only does CMOT outperform existing state-of-art methods, but its inferred gene expression is biologically interpretable by evaluating on emerging single-cell multi-omics datasets. Finally, CMOT is open source at: `https://github.com/daifengwanglab/CMOT`.

2.3 Results

Overview

CMOT (Cross-Modality Optimal Transport) is a computational approach for cross-modality inference of single cells Figure 2.1. CMOT accepts available multimodal and single modality datasets as inputs. CMOT does not require that the available multimodalities have complete corresponding information, i.e., allowing a fraction of unmatched cells in the source.

CMOT first aligns a group of cells X and Y (source) within available multimodal data onto a common latent space (Step A), if the cells across multimodalities do not have complete correspondence. However, this is an optional step if the cells across multimodalities have complete correspondence. In this study, we used Non-linear Manifold Alignment (NMA) [1] to align the unmatched multimodalities. Next, CMOT applies optimal transport to map the cells with a single modality $\widehat{Y}$ (target) to cells in the source from the same modality Y by minimizing their cost of transportation using Wasserstein distance (Step B). This distance can be regularized by prior knowledge (e.g., cell types) or induced cell clusters to improve mapping, and an entropy of transport to speed up OT computations. The optimal transport optimization tries to find an efficient mapping π^* between cells of Y and $\widehat{Y}$ that is used to transport cells Y in to the same space as cells in $\widehat{Y}$. Once transported, CMOT uses k-Nearest-Neighbors to infer the missing modality $\widehat{X}$ for the cells in target $\widehat{Y}$ (Step C). Here, the missing or additional modality $\widehat{X}$ inferred by CMOT has the same number of features as X, and in the same space as X. Details about each step can be found in the Methods section.

We benchmarked CMOT with state-of-art methods [54, 139, 66, 9, 167, 43, 20] on large-scale single-cell multi-omics (e.g., scRNA-seq and scATAC-seq, (Figures A.1- A.2,Figure A.3A.,Figure A.4)). Also, we applied CMOT to additional omics datasets like protein expression. These datasets span across broad contexts including human and mouse brains, cancers and immunology, showing the generalizability of CMOT.

Single-cell gene expression inference from chromatin accessibility in human and mouse brain Human Brain:

We first applied CMOT to single-cell human brain data with jointly profiled chromatin accessibility and gene expression by 10xMultiome (scATAC-seq and scRNA-seq of 8,981 cells) and inferred gene expression of cells from open chromatin regions (OCRs by peaks from scATAC-seq) [149]. We selected the top 1000 most variable genes and peaks (Supplementary Methods). We randomly split the cells into 80% training for cross-validation and 20% testing set for evaluation. We split the training set into training and validation to find optimal parameters for the model using 5-fold cross-validation. For the alignment, we set K=5, and latent dimension d=20. For optimal transport, we set parameters λ=200 and η=1. For KNN modality inference, we set k=600. Also, we used 10 major brain cell types from the dataset. However, to test CMOT's performance when such cell-type information is absent, we induced cell labels by two major cell clusters. We also tested CMOT's performance for different levels of correspondence: p=25%, 50%, 75%, 100%.

CMOT achieves a strong performance for gene expression inference on the testing data, outperforming state-of-the-art methods like Seurat and MOFA+ (Figure 2.2A, 2.2B). For instance, CMOT reports a median cell-wise Pearson correlation r = 0.67 for p=100%, significantly higher than both MOFA+(median r=0.4, Wilcoxon rank-sum test p-value=0) and Seurat (median r=0.64, Wilcoxon p-value<1.23e-

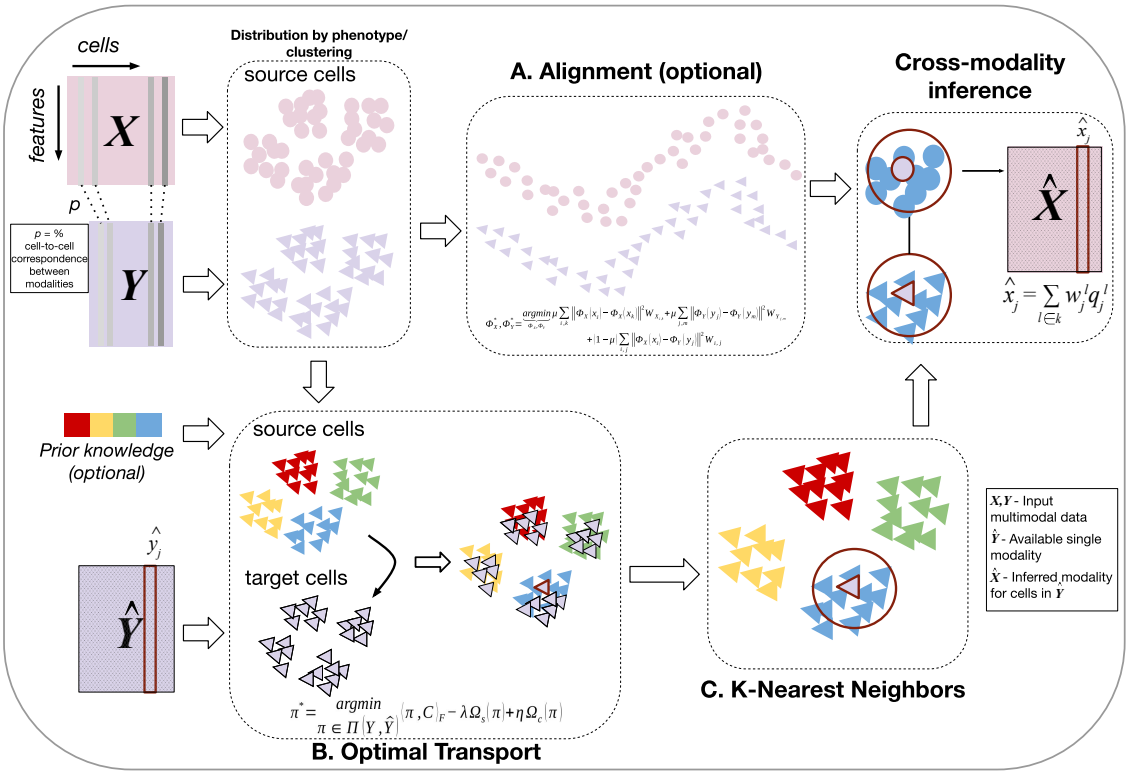

Figure 2.1: **Cross-Modality Optimal Transport (CMOT):** CMOT is a computational approach to infer missing modalities for existing single-cell modalities. It has three main steps: **(A)** Alignment (optional), **(B)** Optimal Transport, and **(C)** k-Nearest Neighbors inference. CMOT inputs two multi-modalities X and Y (source), where the cells in X and Y need not be completely corresponding. The cell-to-cell correspondence information between X and Y can be specified through p. CMOT aligns X and Y onto a common low-dimensional latent space if cells in X, Y do not have complete correspondence. Then, CMOT uses optimal transport (OT) to map the cells in source Y to the cells in target $\hat{Y}$, where Y and $\hat{Y}$ share modalities. CMOT minimizes the cost of transportation by finding the Wasserstein distance between cells in Y and $\hat{Y}$, which is further regularized by prior knowledge or induced cell clusters and entropy of transport. Finally, CMOT infers the missing modality $\hat{X}$ for cells in $\hat{Y}$ using k-Nearest Neighbors (kNN). It calculates a weighted average of the k-nearest mapped cells in Y for every cell in $\hat{Y}$, using their values from X, and infers $\hat{X}$.

14). Even for partial correspondences, CMOT has significantly higher performances (median r=0.65 for p=75%, and r=0.65 for p=50%) than MOFA+ (Wilcoxon p-value<2.8e-294) and Seurat (Wilcoxon p-value<3.43e-10). Also, with low correspondence such as p=25%, CMOT's performance is still significantly higher than MOFA+ (Wilcoxon p-value<1.65e-157). For gene-wise correlation (Figure 2.2B), CMOT p=100% and p=75%, both outperform MOFA+ for 836 versus 118 genes (Wilcoxon p-value<5.38e-118) and 827 versus 140 genes (Wilcoxon p-value<1.78e-165), respectively (Figure 2.2B). Also, CMOT p=100% outperforms Seurat for 494 versus 460 genes (Wilcoxon p-value<2.42e-2). For CMOT p=75%, Seurat slightly performs better for 497 genes versus 471 for CMOT (Wilcoxon p-value<3.01e-1).

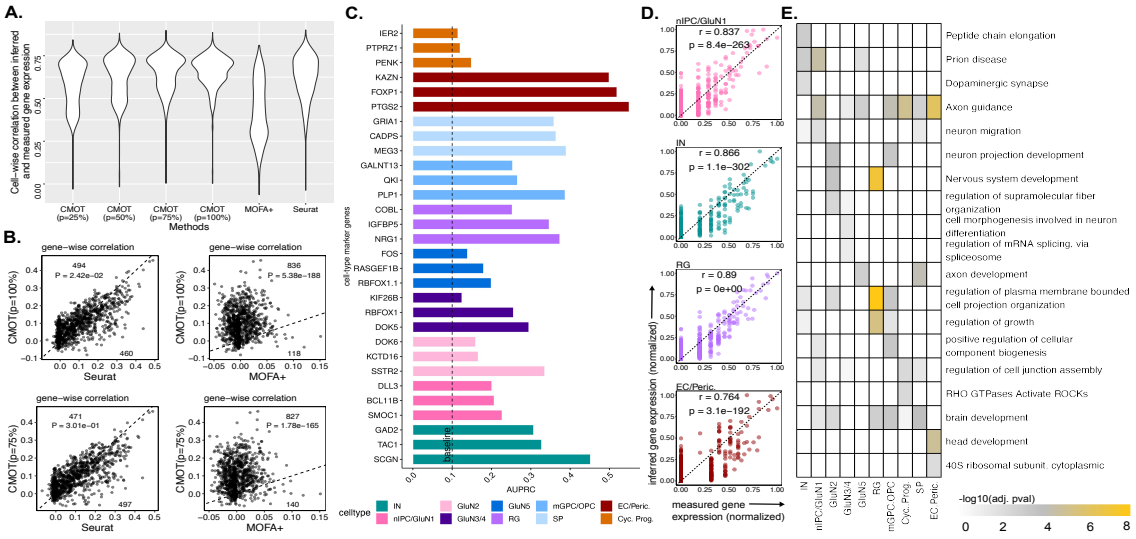

Figure 2.2: **Single-cell gene expression inference from chromatin accessibility in developing human brain (A)** Cell-wise Pearson correlation (y-axis) of inferred and measured gene expression by different methods (x-axis): CMOT (p=25%,50%,75%,100%), Seurat, MOFA+ (Tables A.1, A.2,A.3,A.4, Figure A.1). **(B)** Gene-wise correlation between the inferred and measured gene expression, comparing CMOT (y-axis) with MOFA+ and Seurat (x-axis). Dots: Genes; Numbers: numbers of genes with improved inference by comparing methods. P-values are from one-sided Wilcoxon rank-sum tests. **(C)** Gene-wise AUPRC of cell type marker gene for classifying the cell type. Dashline: baseline=0.1 (see Methods). **(D)** The measured (x-axis) versus inferred normalized expression (y-axis) of genes (dots) for one cell across four cell types: nIPC/GluN1(first), E.C./Peric. (second), IN (third), R.G. (last). (E) Heatmap showing different enriched terms ranked by -log10(adj. p-value of enrichment) values for the top 100 highly predictive genes within each cell type (see Methods). r is Pearson correlation coefficient. p is the correlation p-value.

Next, we evaluated if the CMOT inferred gene expression to classify brain cell types. We used known cell type marker genes provided along with the dataset and selected the top 8 highly predictive cells from each cell type within our inferred gene expression. We then calculated the AUPRC of the respective inferred genes in a one-vs-all manner for each cell type against the rest (Methods). CMOT obtains the higher AUPRCs for these genes against a baseline of 0.1 (Figure 2.2C) for all cell types. The baseline is defined as the proportion of positives in the data. This suggests that the CMOT inferred expression is capable to distinguish cell types, providing the biological interpretability of the CMOT inference. Looking at individual cells (Figure 2.2D), CMOT infers individual cell expression with high Pearson correlation and significance (correlation p-value<0.05). Furthermore, we also found the enriched functions and pathways relating to brain development from the top 100 highly predictive genes (Figure 2.2E). For results of benchmarking on additional state-of-art methods see Figures A.1, Tables A.1, A.2,A.3,A.4, Figure A.1, and Supplemental Methods.

Furthermore, we benchmarked CMOT with the state-of-art methods on a SNARE-seq dataset [26] consisting of jointly profiled gene expression and chromatin peaks

in adult mouse brain. (See Figures A.2, Figures A.3, Tables A.5, A.6,A.7,Tables A.8, A.9,A.10).

Inferring protein expression from gene expression in peripheral blood mononuclear cells

We applied CMOT to infer protein expression from gene expression of peripheral blood mononuclear cells (PBMCs) using emerging CITE-seq data [54]. We trained CMOT on 6885 cells from PBMC10k, with parameters: K=5, d=15, λ=1e−02, η=1, k=100, and used the top 200 highly variable genes in the training data to find the k nearest neighbors. We induced cell labels by identifying two clusters using gene expression for the label regularization in optimal transport. We evaluated CMOT, MOFA+, Seurat, and TotalVI's using 3994 cells from a different dataset, PBMC5k. Here we show an independent evaluation of CMOT and other methods on PBMC5k while using PBMC10k as the training data (Tables A.11, A.12, A.13). Additionally, we also show benchmarking on PBMC10k, by splitting it into 80% training and 20% testing data (see Figures A.7, Figures A.14, Tables A.15, A.16).

As shown in Figure 2.3A, CMOT achieves a median cell-wise Pearson correlation=0.86 for p=100%, significantly outperforming MOFA+ (Wilcoxon p-value<6.9e-57) and TotalVI (default parameters) (Wilcoxon p-value=0) as well as comparable with Seurat. For instance, we show two cells and their Pearson correlation of inferred versus measured protein expressions (r=0.99, p=1.4e-11 and r=0.98, p=6.1e-11) in Figure 2.3B. Moreover, even for partial correspondences, p=25%, 50%, 75%, CMOT performs consistently with significantly higher cell-wise correlation than MOFA+ (Wilcoxon p-values=0,0,0) and TotalVI (Wilcoxon p-values<8.36e-58, 1.73e-45, 5.25e-12). Also, for inferring individual protein expression, CMOT has high correlations for all proteins, consistent with state-of-art methods (Figure 2.3C), with some examples shown in Figure 2.3B. Rest of the proteins along with their inference statistics are reported in Figures A.5 and Tables A.11, A.12,A.13.

Inference of gene expression using chromatin accessibility for drug-treated lung cancer cells

Next, we applied CMOT to 100nM dexamethasone (DEX)-treated A549 single-cells from lung adenocarcinoma. The cells were profiled after 0, 1, and 3 h of treatment for gene expression and open chromatin regions (OCRs) using sci-CAR experiments [17]. We focus on the CMOT's performance for gene expression inference from peak signals of OCRs. We stratified-split the dataset into 80% training and 20% test cells using the treatment hours. We used the treatment hours as the classes for label regularization in optimal transport for training cells. We trained CMOT with the parameters: K=5, d=10, λ=1e02, η=5e-3, k=500, and used the top 20 highly variable OCRs in scATAC-seq to find the k nearest neighbors. Again, we found that CMOT shows a consistent performance across different cell-to-cell correspondence information (p) with high correlation. CMOT (p=100%) infers gene expression

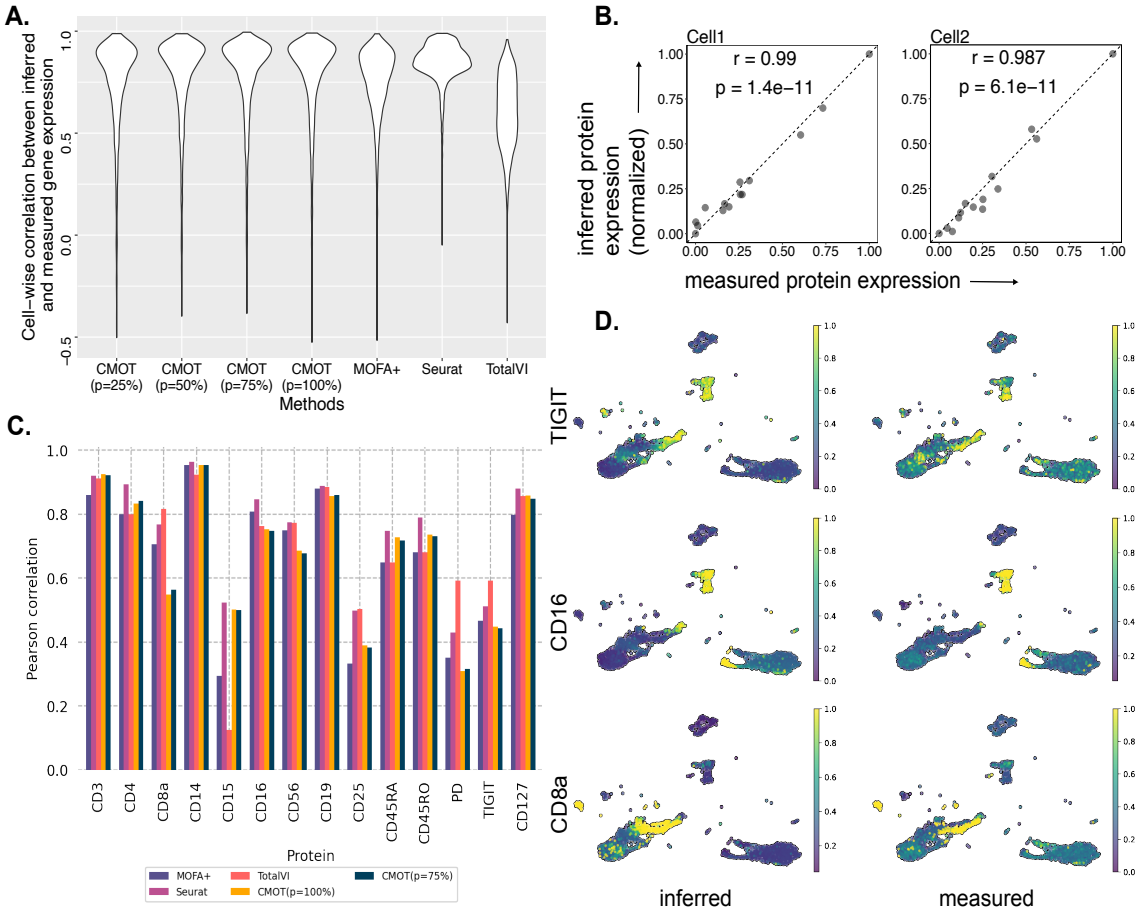

Figure 2.3: **Inferring protein expression from gene expression in single-cell peripheral blood mononuclear cells (A)** Cell-wise Pearson correlation (y-axis) of inferred and measured protein expression by different methods (x-axis): CMOT(p=25%,50%,75%,100%), Seurat, MOFA+, TotalVI (Tables A.11, A.12,). **(B)**The measured (x-axis) versus inferred normalized expression (y-axis) of 14 proteins (dots) for two select cells. **(C)** Pearson correlations of inferred and measured expression (y-axis) of individual proteins (x-axis) by CMOT(p=100%,75%), Seurat, MOFA+, TotalVI (Tables A.13). **(D)**UMAPs of inferred and measured expressions for three proteins: TIGIT (r=0.45; p=6.18e-197) (top), CD16 (r=0.75; p=0) (middle), CD8a (r= 0.55; p=3.11e-312) (bottom) (Tables A.13). The intensity represents the protein expression level. r is Pearson correlation coefficient. p is the correlation p-value.

with a median Pearson correlation of 0.52, similar to MOFA+ and significantly outperforming Seurat (median correlation=0.41, Wilcoxon p-value<2.19e-57) (Figure 2.4A). Moreover, CMOT shows a high gene-wise Pearson correlation outperforming Seurat for 636 versus 547 genes (Figure 2.4B, Tables A.21). Although MOFA+ reports a higher gene-wise Pearson correlation for some genes than CMOT (Figure 2.4B, Tables A.21), we still see that CMOT's inferred expression shows the transitory trend of key druggable marker genes across drug-treatment hours. Figure 2.4C shows three key genes, identified as makers of early (ZSWIM6) [103] and late (PER1, BIRC3) [125], [14], [124] events of treatment. Also, we performed enrichment of the

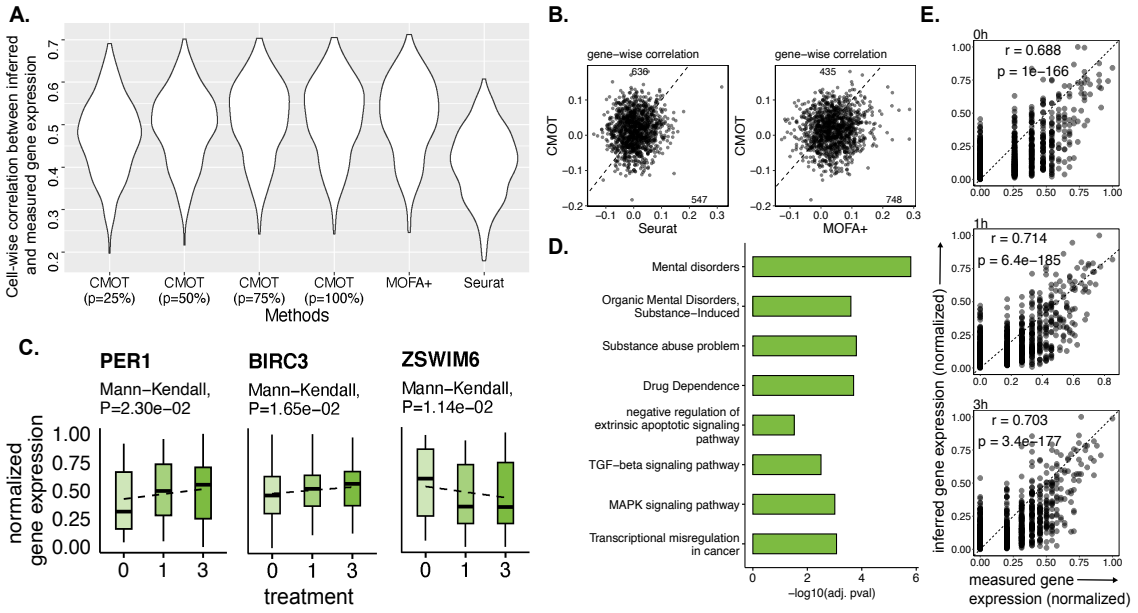

Figure 2.4: **Inference of gene expression for drug-treated lung cancer cells using chromatin accessibility (A)** Cell-wise Pearson correlation (y-axis) of inferred and measured gene expression by different methods (x-axis): CMOT(p=25%,50%,75%,100%), Seurat, MOFA+, TotalVI (Tables A.17, A.18,A.19,A.20). **(B)** Gene-wise correlation between the inferred (y-axis) and measured (x-axis) expression, comparing CMOT with MOFA+ and Seurat. Dots: Genes; Numbers: Gene numbers above and below the dotted line. P-values are calculated by a one-sided Wilcoxon rank-sum test (Tables A.21,). **(C)** CMOT inferred normalized gene expression trend (y-axis) across treatment hours (x-axis). Key genes: PER1 and BIRC3 [125], [14], [124] are markers for glucocorticoid receptor (GR) activation seen later in treatment (3h). ZSWIM6 [103] is a key gene of early events of DEX treatment (0h,1h). **(D)** Enriched terms associated with CMOT inferred gene expression using 435 genes with a higher gene-wise Pearson correlation compared to MOFA+'s 748 genes inference in (**B**, Tables A.6B,)(**E**) The measured (x-axis) versus inferred normalized expression (y-axis) of genes (dots) for three select cells. r is the Pearson correlation coefficient. p is the correlation p-value.

435 high correlation genes identified by CMOT (versus MOFA+ in **Fig. 4B**). As shown in Figure 2.4D, we saw a higher enrichment of terms associated with DEX-treated A549 cells like TGF-beta signaling, along with effects on DEX treatment in general, like Mental disorders as compared to enrichment given by MOFA+ (Figure A.6). Lastly, we also found that the cell-wise correlations between inferred and measured gene expression are also significantly highly correlated in each treatment hour (Figure 2.4E). For results of benchmarking on additional state-of-art methods see Figures A.4 and Tables A.17, A.18,A.19,A.20.

Cross-modality inference between gene expression and chromatin accessibility to distinguish cancer types

Finally, we tested CMOT to see how well it can infer between two modalities, especially for relevantly small datasets. We used a pan-cancer scCAT-seq dataset which jointly profiled single-cell gene expression and chromatin accessibility on OCRs for

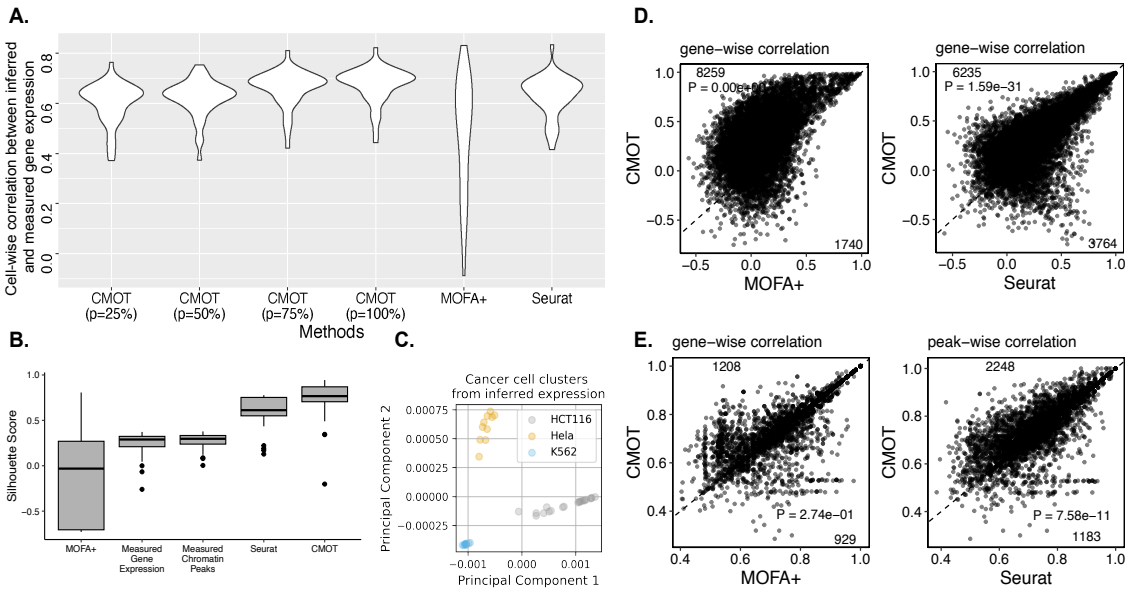

Figure 2.5: **Cross-modality inference between gene expression and chromatin accessibility can distinguish cancer types (A)** Cell-wise Pearson correlation (y-axis) of inferred and measured gene expression by different methods (x-axis): CMOT(p=25%,50%,75%,100%), Seurat, MOFA+, TotalVI (Tables A.22, A.23,A.24,A.25). **(B)**Silhouette score (x-axis) across measured and inferred gene expressions (x-axis) (Table A.26). **(C)** PCA of inferred gene expression. **(D)** Gene-wise correlation between the inferred and measured expression, comparing CMOT (y-axis) with MOFA+ and Seurat (x-axis). Dots: Genes; Numbers: Gene numbers above and below the dotted line. (**E**)Peak-wise AUROC, comparing CMOT (y-axis) with MOFA+ and Seurat (x-axis). Dots: Peaks; Numbers: Peak numbers above and below the dotted line. P-values are calculated by a one-sided Wilcoxon rank-sum test.

three cancer cell lines: HCT116, HeLa-S3, and K562 [101]. We stratified split data into 60% training and 40% testing sets using cancer-type information. We induced our own cell labels for training cells for label regularization in optimal transport. For gene expression inference from OCR peaks, we identified two clusters in chromatin peaks and vice versa. We trained CMOT with the following parameters for gene expression inference from chromatin peaks: K=5, d=10, λ=1e03, η=1, k=40, and used the top 150 highly variable OCRs to find the k nearest neighbors. For inferring gene expression from binarized OCR peaks, we evaluated the inferred expression using the same metrics (cell-wise and gene-wise Pearson correlation) as above.

CMOT significantly outperforms both MOFA+ and Seurat, with a cell-wise median correlation of 0.69 compared to 0.65 (Wilcoxon p-value<6.81e-17) and 0.51 (Wilcoxon p-value<1.32e-05), respectively (Figure 2.5A, Tables A.22, A.23,A.24,A.25). Moreover, CMOT (p=100%) yields an improved gene-wise correlation for 6307 genes versus 3693 against Seurat (Wilcoxon p-value<7.19e-18), and, 8734 versus 1266 against MOFA+ (Wilcoxon p-value=0) (**Fig. 5D**). Moreover, CMOT's inference is particularly useful to identify the cancer type specific cell clusters. For instance, we calculated the Silhouette score (see Methods) to see if the cells from the same cancer

lines exhibit similar gene expression patterns. CMOT reports a high median silhouette score of 0.82 compared to the measured gene expression (0.35) and inferred expressions from Seurat (0.75) and MOFA+ (0.05) (Figure 2.5B, Table A.26). As shown in Figure 2.5C, the cancer cells from three cancer cell lines can be separated using CMOT-inferred gene expression, suggesting the capability of CMOT inference to reveal cancer-type-specific expression.

Then, we evaluated the CMOT's OCR peaks inference from gene expression. We trained CMOT with the parameters: K=5, d=10, λ=1e03, η=1, k=10, and used the top 50 highly variable genes to find the k nearest neighbors. We also stratified split the data into 60% training and 40% testing sets using cancer-type information. We normalized CMOT's inferred peaks and then binarized them by a cutoff 0.5, and then calculated the peak-wise area under the receiver operating curve (AUORC) of the inferred binarized peaks relative to the binarized measured profile. We also found that CMOT significantly outperforms both MOFA+ and Seurat with Wilcoxon p-values<2.71e-86,2.70e-72 respectively for OCR peak inference from gene expression (Figure 2.5E, Figures A.8 and Tables A.27, A.28).

2.4 Methods

Overview of Cross Modality Optimal Transport

CMOT (Cross-Modality Optimal Transport) is a computational approach for cross-modality inference of single cells. As shown in Figure 2.1, CMOT has three main steps: **Step A** – Alignment to project the cells with available multimodal data (source cells) onto a common low-dimensional latent space. However, this is an optional step if the cells across multimodalities have complete correspondence; **Step B** – Optimal Transport to map the cells with a single modality (target cells) to the aligned source cells from the same modality. To this end, we minimize the Wasserstein distance between the source and target cells and can further regularize the minimization using prior knowledge (e.g., cell types) or induced cell clusters and entropy of transport; **Step C** – k-Nearest-Neighbors to infer the missing or unprofiled modality of target cells using another modality of nearest mapped source cells.

Step A: Alignment of source cells with multimodal data

We set this as an optional step if the cells across available multimodalities do not have a complete correspondence. That is, if cells across modalities have none to partial correspondence between them, then CMOT first aligns the. Although users are free to use their choice of alignment in such scenarios, we use Nonlinear Manifold Alignment (NMA) [1].

Alignment is an important step that accounts for when the source cells have partial correspondence. NMA is based on a manifold hypothesis that high dimensional multimodal datasets have similar underlying low dimensional manifolds, and therefore, they can be projected onto a common manifold space that preserves the local

geometry of each modality and minimizes the differences between the manifolds of modalities. We define $X = \{x_i\}_{i=1,..,s_X}$ and $Y = \{y_j\}_{j=1,..,s_Y}$ as two multimodal measurements with s_X, s_Y source cells in Modalities X, Y respectively, where $x_i \in R^{dX}$ and $y_j \in R^{d_Y}$ represent the measurements of d_X features in i^{th} cell of Modality X, and d_Y features in j^{th} cell of Modality Y. We also define $W_X \in R^{s_X \times s_X}$ and $W_Y \in R^{s_Y \times s_Y}$ as cell similarity matrices for X and Y, respectively, where each similarity matrix is constructed by connecting a cell with its K nearest neighboring cells within the modality. We define p $(0 < p < 100\%)$ to quantify the partial prior known cell-to-cell correspondence information (for example, $p\%$ of paired cells across modalities) and encode this information as a cross-modal similarity matrix $W \in R^{s_X, s_Y}$. NMA then learns two mapping functions ϕ_X and ϕ_Y that project x_i and y_j to $\phi_X(x_i) \in R^d$ and $\phi_Y(y_j) \in R^d$, respectively onto a common manifold space with dimension $d \ll min(d_X, d_Y)$. The d-dimensional manifold preserves the local geometry of each modality and minimizes the distances between corresponding samples after projection. Solving manifold alignment can be reformulated as manifold co-regularization in reproducing kernel Hilbert spaces. The manifold alignment optimization finds optimal mapping functions ϕ_X^*, ϕ_Y^* by solving the following:

$$\phi_X^*, \phi_Y^* = \underset{\phi_X, \phi_Y}{\arg\min}\, \mu \sum_{i,k} ||\phi_X(x_i) - \phi_X(x_k)||^2 W_{X_{i,k}} + \mu \sum_{j,m} ||\phi_Y(y_j) - \phi_Y(y_m)||^2 W_{Y_{j,m}} + (1-\mu) \sum_{i,j} ||\phi_X(x_i) - \phi_Y(y_j)||^2 W_{i,j} \tag{2.1}$$

, where the first two terms preserve the local geometry within each modality, the similarity matrices W_X and W_Y model the relationships of the cells in each modality that can be identified by the k-nearest neighbor graph, and the third term preserves the correspondence information across X and Y modeled by W. The parameter μ controls the trade-off between conserving the local geometry of each modality and cell-to-cell correspondences across modalities. Here, we set μ to 0.5.

We also need to add an additional non-zero constraint to avoid mapping of all cells onto a latent space with dimension zero:

$$\boldsymbol{P}^T D\, \boldsymbol{P} = I, \quad \boldsymbol{P} = \begin{bmatrix} \phi_X \\ \phi_Y \end{bmatrix},$$

where

$$\boldsymbol{\phi}_X = [\![\, \phi_X(x^1), \ldots, \phi_X(x^{s_X}) \,]\!]^d, \quad \boldsymbol{\phi}_Y = [\![\, \phi_Y(y^1), \ldots, \phi_Y(y^{s_Y}) \,]\!]^d,$$

$\boldsymbol{D}$ is the diagonal matrix of $\mu W_X, \mu W_Y$, and I is the identity matrix [114].

Also, two modalities are not required to have complete correspondence between the cells. Therefore, W is a binary correspondence matrix between cells of X and Y such that if p=100, i.e., 100% correspondence across cells in X and Y, W would be an identity matrix. For $p < 100$, $W_{i,j} = 1$ if i^{th} and j^{th} cells from Modalities

X and Y respectively are the corresponding cells and 0 otherwise. After alignment, the resulting d-dimensional modalities share a common latent space that can easily be compared using Euclidean distances. For instance, for every cell $y_j \in Y$, we find an aligned cell $x_{j,a} \in X$ by finding the closest cell in X using the Euclidean distance. To implement our alignment step, we used the non-linear manifold module from our published Python package ManiNetCluster [114].

Unless otherwise stated, we use the term CMOT for our model trained with full correspondence ($p = 100\%$).

Step B: Optimal Transport to map source and target cells by shared modality

The optimal transport theory [22], [83] tries to find the most efficient mapping π^* that transports one probability distribution to another with minimum transportation cost. A mapping $\pi \in \Pi(Y, \hat{Y})$ represents transport plan to map cells from source (Y) and target ($\hat{Y}$) modalities, where $\Pi(Y, \hat{Y})$ contains all probabilistic mappings between the source (Y) and target ($\hat{Y}$) cells. We define $\hat{Y} = \{\hat{y_j}\}_{j=1,..,s_{\hat{Y}}}$ as the target single modality measurement with $s_{\hat{Y}}$ cells, where $\hat{y}_j \in R^{d_Y}$ represents the measurement of d_Y features in j^{th} cell of Modality $\hat{Y}$. The classical OT distance (Wasserstein distance) gives mappings between two probability distributions as the transportation cost. Let C be the cost matrix where $C \geq 0 \in R^{s_X, s_y}$ and $C_{i,j}$ represents the pairwise cost of mapping the source cell $y_i \in Y$ to the target cell $\hat{y}_j \in \hat{Y}$. For discrete probability distributions like Y and $\hat{Y}$ over the same metric spaces (i.e., matched features of the shared modality), we define the OT problem as:

$$\pi^* = \min_{\pi \in \Pi(Y, \hat{Y})} \langle \pi, C \rangle_F - \lambda \Omega_s(\pi) + \eta \Omega_c(\pi), \tag{2.2}$$

where the first term computes the Frobenius dot product $\langle ., . \rangle_F$ between the cost matrix C and π. The set Π is defined as $\Pi(Y, \hat{Y}) = \{\pi \in R_+^{s_Y, s_{\hat{Y}}} : \pi \mathbb{1}_{s_{\hat{Y}}} = Y, \pi^T \mathbb{1}_{s_Y} = \hat{Y}\}$.

The second term, also called entropic regularization, calculates the entropy of transportation for π where $\Omega_s(\pi) = -\sum_{i,j} \log \pi(i, j)$. Entropic regularization addresses the computational complexity of OT as the sample size increases [32]. The intuition behind this term is to relax the sparsity constraints of the OT problem by increasing its entropy so that π^* is denser, as source cells (Y) are distributed more towards target cells ($\hat{Y}$). The resulting formulation is strictly convex and can be solved through Sinkhorn's Algorithm [34]. The parameter λ weights the entropic regularization. As the parameter increases, the sparsity of π^* decreases, giving a smoother transport.

The third term is the label regularizer [32], $\Omega_c = \sum_j \sum_c ||\pi(I_c, j)||_p^q$, where I_c contains the index of rows in π related to the source cells (Y) that belong to class c if we have such prior knowledge, e.g., known cell types. Here, $\pi(I_c, j)$ is a vector containing the coefficients of the j^{th} column of π associated with class c, where the j^{th} column in $\pi(I_c, j)$ represents the j^{th} target cell in $\hat{Y}$. The norm $._p^q$ denotes the l_p norm to the power of q. The parameter η weights the label regularization. The

intuition behind this term is to penalize the mappings that match together samples from different labels. This means that even if we do not have the label information for the target cells ($\hat{Y}$), we can promote group sparsity within the columns of π such that each target cell is only associated with a class. However, in the absence of such label information, we can compute our own labels through unsupervised clustering techniques like hierarchical clustering to induce cell clusters as labels for source cells in Y. Finally, to map the source cells (Y) to the target space ($\hat{Y}$), we use barycentric mapping $Y^{(t)} = \pi^* Y$ [32]. Now, we can easily compare $Y^{(t)}$ and $\hat{Y}$ using euclidean distance. To solve the regularized OT optimization step, we used Domain Adaptation functions (*ot.da*) provided in the Python package Python optimal transport (POT) [130].

Step C: k-Nearest Neighbors to infer the missing modality of target cells

Finally, we apply k-Nearest Neighbors (kNNs) to infer the missing modality $\hat{X}$ of target cells in $\hat{Y}$. For each target cell $\hat{y_j} \in \hat{Y}$, we find its kNN in $Y^{(t)}$ using Euclidean distance. Let $S_j = \{c_j^l : l = 1, 2, .., k\}$ be the set of k nearest neighboring cells of $\hat{y_j}$ in $Y^{(t)}$, where c_j^l is a cell in $Y^{(t)}$. For cells in S_j, we use their values from the aligned modality X to define another set $Q_j = \{q_j^l : l = 1, 2, .., k\}$, where q_j^l represents the profile of the cell c_j^l within the aligned modality X. Finally we calculate the weighted average of the profiles of all cells in Q_j to get $\hat{x_j}$. This is calculated as:

$$\hat{x_j} = \sum_{l \in k} w_j^l q_j^l \tag{2.3}$$

, where w_j^l is the weightage given to q_j^l using $w_j = \exp\left(-\sqrt{||\hat{y_j} - y_{S_j}||^2}\right)$. Thus, we get the corresponding modality $\hat{X}$ for $\hat{Y}$. We used sklearn's [45] nearest neighbor function for kNN implementation.

Single-cell multi-omics datasets

We tested CMOT on four single-cell multiomics datasets: (1) Gene expression and chromatin accessibility of single cells in human and mouse brains (scRNA-seq and scATAC-seq) [149], [26]; (2) Gene and protein expression of peripheral blood mononuclear cells (CITE-seq) [54]; (3) Gene expression and chromatin accessibility of A549 lung cancer cells (sci-CAR) [17]; (4) Gene expression and chromatin accessibility of pan-cancer cells (scCAT-seq) [101]. All details on data and data processing are available in supplementary methods.

Partial correspondence in multi-omics data:

Joint profiling of single cells is challenging and therefore, it may not always be feasible to get completely corresponding cells across profiled modalities. In such scenarios, there could be partial to no correspondence across cells of multimodalities. For example, a 50% cell-to-cell correspondence between modalities means that only 50% of the cells have been jointly profiled between the modalities. As a result,

training on partially corresponding multimodalities for cross-modality inference can lead to misleading or wrong inferences. Therefore, to address this problem, CMOT first aligns such partially corresponding datasets, and then performs inference. In this paper, we have used datasets that have a 100% correspondence originally, so that we can validate the inference performance. However, we simulate different levels of cell-to-cell correspondence by setting the p value in non-linear manifold alignment (Methods Step A). We report CMOT's performance when trained on $p = 25\%, 50\%, 75\%, 100\%$ cell-to-cell correspondence levels, and show that CMOT's cross-modality inference performance can beat state-of-the-art methods that require 100% cell-to-cell correspondence for training.

Runtime evaluations

We compare CMOT's running time with state-of-art methods MOFA+, Seurat, TotalVI, and Polarbear for the best-performing parameters used for cross-modality inference (Table A.29). We benchmarked all methods on Intel Xeon Gold 6242R CPU @3.10GHz x 40 with 251.4GiB RAM and NVIDIA RTX A6000 GPU. We induced cell labels for datasets with no prior knowledge (e.g., cell types). We use these labels for label regularization in OT optimization (see Methods and Materials Step B) to improve the mappings between cells in the source (X) and target (Y) modalities. To induce cell labels, we performed hierarchical clustering of training and validation sets combined using the scikit-learn clustering functions[118].

Training and cross-validation

We split the human brain [149], PBMCs [54], DEX-A549 lung cancer [17], and pan-cancer [101] datasets into 80% train and 20% test. We trained all methods: Seurat [139],[66], MOFA+ [8], TotalVI [54], and Polarbear [167] using default parameters for all datasets except DEX-treated A549 [17]. For modality inference in Seurat, we integrated the training modalities first, then we inferred the missing modality using FindTransferAnchor and TransferData functions [139] between the integrated training modalities and source test modality. For MOFA+, we input the missing modality as NA values and trained the model on the multimodalities. We trained TotalVI autoencoder with default parameters, with latent distribution set to "normal", on the training set. Finally, we trained Polarbear and Polarbear co-assay models using default parameters on the training set. To identify the highest performing parameters for Steps B and C of CMOT, we performed 5-fold cross-validation on the training set. We reported the best-performing parameters for each dataset in the Results. For the DEX-treated A549 dataset, we tuned parameters for all methods (Figure A.9) and benchmark inference performance of CMOT against state-of-arts (Fig. 4A, Figure A.4).

Parameter selection

In Step A, we found the optimal alignment by testing different values of d common manifold dimensions and K nearest neighbors for building the similarity within each modality (Figure A.10). In Step B and Step C, we performed cross-validation to select the regularization coefficients λ and η for optimal transport and k-nearest neighbors based on CMOT's performance saturation (Figure A.11). Also, for all datasets applied in this paper, we held out a 20% testing set to report CMOT's performance. For datasets with no prior knowledge (e.g., cell types), we induced cell labels by cell clusters through hierarchical clustering of the training set, when training the final model. We split the training data into training and validation sets to select parameters through 5-fold cross-validation (see Supplementary Methods).

Evaluation

Inference versus measurement:

To evaluate CMOT's inferred gene and protein expressions, we calculated Pearson's correlation coefficient between the inferred and measured expression values of each cell (cell-wise). Also, we computed the gene-wise correlation between inferred and measured expression values across cells for each gene [167]. For peak inference in open chromatin regions, we used AUROC to evaluate the quality of CMOT's binarized inferred peaks [167]. We computed peak-wise AUROC between individual inferred peak profiles versus measured profiles. This evaluation also applied to the state-of-art methods that we compared. We reported the number of genes with improved correlation/AUROC w.r.t. these methods along with a one-sided Wilcox rank-sum test p-value for each [167].

Classifying known cell type using inferred expression:

For the human brain data with known brain cell type information, we evaluated the CMOT inferred expression of cell-type marker genes for classifying the cell type and calculated the AUPRC of the classification [167]. To this end, given a cell type, we labeled all cells that belong to the cell type as positive and the rest as negative. Specifically, we evaluated the Top 8 marker genes from each cell type, due to disproportionate cell-type distribution within the dataset, using a total of 80 cells. We then defined a baseline = 0.1 for the AUPRC as the ratio of the number of positives versus total cells.

Clustering cancer types using inferred gene expression by Silhouette Score:

For the pan-cancer dataset, we evaluated CMOT to separate the cancer types. In particular, we assessed if CMOT's inferred gene expression data can cluster the cells and cell clusters correspond to different cancer types [101], using the silhouette score. The silhouette score $S(m)$ of a cell m belonging to cluster C_M is calculated as:

$$S(m) = \frac{E(m) - e(m)}{\max(E(m), e(m))} \tag{2.4}$$

, where $E(m) = \min_{M \neq N} \frac{\sum_{n \in C_N} d(m,n)}{|C_N|}$ is the inter-cluster distance defined as the average distance to closest cluster of cell m except that which it's a part of (i.e. C_N) and $e(m) = \frac{1}{|C_M|-1} \sum_{n \in C_M, m \neq n} d(m,n)$ is the intra-cluster distance defined as the average distance to all other cells in the cluster to which it's a part of. We calculated the Silhouette scores by the Python package Scikit-learn [45].

Gene set enrichment analysis:

We used Metascape [168] to perform gene set enrichment analysis for the highly predictive genes by CMOT.

Comparison with state-of-the-arts:

We compared CMOT with existing state-of-the-art methods, Seurat [66], MOFA+[8], TotalVI [54], and Polarbear [167].

First, for the human brain data [149], we benchmarked CMOT against Seurat and MOFA+ for the human brain data (Figure 2.1,Figure A.1); for mouse brain [26] we only compare CMOT to Polarbear because Polarbear has previously been applied to the same data (Figure A.2, Figure A.3). However, when we tried running Polarbear on other datasets, we found it difficult to run due to several dataset-specific hard coded variables used in the model.

Next, for the CITE-seq data [54], we compared CMOT with Seurat, MOFA+, and, TotalVI (Figure 2.3,Figure A.7). We added TotalVI to the comparison since it was specifically designed for CITE-seq datasets.

For the DEX-treated A549 dataset [17], we benchmarked CMOT against Seurat and MOFA+ (Figure 2.4), as well as other state-of-art methods (Figure A.4).

For the pan-cancer dataset [101], we benchmarked Seurat and MOFA+ due to the small dataset size (Figure 2.5).

Chapter 3

Modelling cell trajectories, population changes and perturbation effects in time-series single-cell transcriptomics

3.1 Abstract

Cellular processes like development, differentiation, and disease progression are highly complex and dynamic (e.g., gene expression). These processes often undergo cell population changes driven by cell birth, proliferation, and death. Single-cell sequencing enables gene expression measurement at the cellular resolution, allowing us to decipher cellular and molecular dynamics underlying these processes. However, the high costs and destructive nature of sequencing restrict observations to snapshots of unaligned cells at discrete timepoints, limiting our understanding of these processes and complicating the reconstruction of cellular trajectories. To address this challenge, we propose ARTEMIS, a generative model integrating a variational autoencoder (VAE) with unbalanced Diffusion Schrödinger Bridge (uDSB) to model cellular processes by reconstructing cellular trajectories, reveal gene expression dynamics, and recover cell population changes. The VAE maps input time-series single-cell data to a continuous latent space, where trajectories are reconstructed by solving the Schrödinger bridge problem using forward-backward stochastic differential equations (SDEs). A drift function in the SDEs captures deterministic gene expression trends. An additional neural network estimates time-varying kill rates for single cells along trajectories, enabling recovery of cell population changes. Using three scRNA-seq datasets—pancreatic β-cell differentiation, zebrafish embryogenesis, and epithelial-mesenchymal transition (EMT) in cancer cells—we demonstrate that ARTEMIS: (i) outperforms state-of-art methods to predict held-out timepoints, (ii) recovers relative cell population changes over time, and (iii) identifies "drift" genes driving deterministic expression trends in cell trajectories. Furthermore, *in silico* perturbations show that these genes influence processes like EMT. The code for ARTEMIS: `https://github.com/daifengwanglab/ARTEMIS`.

3.2 Introduction

Cellular processes such as differentiation (e.g., stem cell differentiation into pancreatic β-cells [152]), development (e.g., embryogenesis in zebrafish [46]), and disease progression (e.g., epithelial-mesenchymal transition (EMT) in A549 lung cancer cells [31]) are dynamic and highly complex. The advent of single-cell sequencing technologies has enabled the capture of gene expression at cellular resolution. However, these experiments are often expensive and destructive, yielding only snapshots of unaligned cells at discrete timepoints. This limitation impedes the analysis of cellular processes, as continuous gene expression across timepoints is unavailable, and reconstructing dynamic trajectories without cell lineage tracking remains challenging. Moreover, cellular processes typically involve continuous population changes, due to cell birth, proliferation, and death (Figure 3.1a.). A model that reconstructs cellular trajectories from discrete timepoints while accounting for population changes and identifying key genes driving cellular processes is crucial (Figure 3.1b.).

Numerous methods have been developed for trajectory inference in single-cell gene expression data. Pseudotime analysis orders cells along trajectories based on gene expression similarity [62, 148, 161]. RNA velocity methods predict future transcriptional states by analyzing unspliced and spliced mRNA ratios in scRNA-seq data [55, 123]. However, both approaches rely on single snapshots, lacking temporal measurements, and thus limiting their ability to fully capture the dynamics of cellular processes.

Dynamic model-based approaches aim to address these limitations by learning continuous trajectories in real time. For example, PRESCIENT employs stochastic differential equations (SDEs) to model cellular differentiation, incorporating both deterministic and stochastic effects [163]. An SDE is formulated as $\mathrm{d}x_t = b(x)\mathrm{d}t + \varepsilon \mathrm{d}W_t$, where the drift term $b(x)$ reflects deterministic trends derived from an energy potential's gradient. PRESCIENT uses neural networks to learn this energy landscape. PI-SDE extends this approach with a physics-informed Hamilton-Jacobi (HJ) loss to improve training stability and identify least-action paths [81]. While these methods use growth rates derived from prior knowledge, they do not model population changes from cell birth, proliferation, or death events.

Optimal transport-based methods, like Waddington-OT, use unbalanced OT to infer probabilistic couplings between successive timepoints [133], but these approaches learn static, linear mappings. TrajectoryNet combines dynamic OT and continuous normalized flows, using neural ordinary differential equations (ODEs) to model cell population dynamics over time [145]. MIOFlow improves upon this by combining SDEs with a geodesic autoencoder, capturing non-linear latent spaces and preserving cellular variations [77]. However, these methods often struggle to generalize to unmeasured timepoints with distinct distributions due to fixed latent spaces. scNODE addresses this by combining a VAE with neural ODEs, incorporating dynamic regularization to enhance robustness against distributional shifts [166]. Nonetheless, ODE-based methods remain limited in capturing the stochasticity in-

herent in cellular processes.

Schrödinger bridges (SBs), a framework for dynamic entropy-regularized OT, have been applied to various fields including generative modelling [27, 37, 156], sampling [12, 74], and biological processes [72, 117]. SBs identify the most likely stochastic evolution between two probability distributions, given a prior or reference process (e.g., Brownian motion). Diffusion Schrödinger bridges (DSBs) [37], building on recent developments in diffusion models [136, 70] and forward-backward SDE theory [27], model both drift and diffusion components of continuous processes such as gene expression. Unbalanced DSBs (uDSBs) extend this framework to unnormalized distributions, allowing for changes in population size, including cell death and proliferation [117]. In contrast, neural ODEs lack diffusion terms and typically assume fixed populations. While uDSBs offer flexibility for modeling cellular dynamics, applying them to high-dimensional data like scRNA-seq remains challenging due to the curse of dimensionality [86].

Here, we present ARTEMIS (trAjectory infeRence wiTh unbalancEd dynaMic optImal tranSport), a generative model that integrates VAE with uDSBs to learn continuous gene expression dynamics and cell populations changes (Figure 3.1c.). ARTEMIS first pre-trains a VAE to map scRNA-seq data into a low-dimensional latent space. The uDSB learns cellular trajectories by solving the SB problem through forward-backward SDEs, learning optimal forward-backward drift functions in this latent space. The VAE and uDSB are jointly trained to ensure a smooth latent space to predict cellular trajectories. Additionally, a neural network predicts time-varying kill rates, which are further used to infer cell statuses (e.g., birth, proliferation, death) along trajectories.

We benchmark ARTEMIS on three time-series scRNA-seq datasets and compare its performance to state-of-the-art methods, including PRESCIENT, MIOFlow, and scNODE, as well as using uDSB as a baseline. Our results demonstrate that ARTEMIS: i) accurately predicts single-cell gene expression at held-out timepoints, ii) recovers relative cell population changes over time, iii) learns a drift (Q) function in the SDE which captures deterministic trends in gene expression dynamics and identifies drift-genes along cellular trajectories. Furthermore, ARTEMIS enables the modeling of *in silico* perturbations introduced at intermediate timepoints (Figure 3.1b.).

3.3 Materials and methods

Overview

ARTEMIS takes time-series scRNA-seq data for measured timepoints $t \in \{0, 1, 2, .., T\}$ as input, and reconstructs the cellular trajectory from $t = 0$ to $t = T$. This includes gene expression for unmeasured timepoints $\tilde{t}$, modeling relative cell population changes, and identifying genes associated with cell drift driving the trajectory. First, it pre-trains a VAE on single-cell gene expression $X_t \in \mathbb{R}^{n_t \times g}$ for measured timepoints t, given n_t cells and g genes by minimizing the L_{vae} loss (see 3.3). The

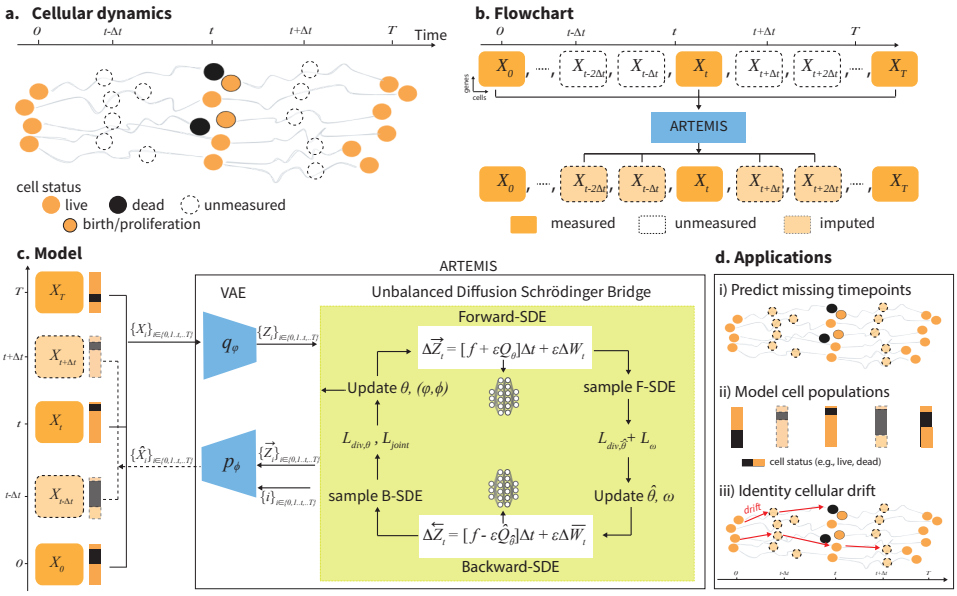

Figure 3.1: **Model overview.** a.) Cellular processes are complex and dynamic, and undergo cell population changes driven by birth, proliferation, and death with time. b.) Single-cell sequencing provides snapshots of unaligned cells at discrete timepoins. To reconstruct cellular trajectories, we propose ARTEMIS. c.) ARTEMIS leverages single-cell time-series gene expression data. It integrates and jointly trains a Variational Autoencoder (VAE) and unbalanced diffusion Schrödinger Bridge (uDSB) to learn a smooth latent space. The uDSB solves the SB problem using forward and backward SDEs, learning optimal backward ($\widehat{Q}_{\widehat{\theta}}$) and forward drifts ($Q_\theta$), along with VAE parameters φ,ϕ by optimizing $L_{div,\widehat{\theta}}$ $L_{div,\theta}$, and, L_{joint}, respectively. To further learn cell population changes, an additional loss L_ω is optimized. d.) ARTEMIS allows following downstream analysis i) predict gene expression for unmeasured timepoints, ii) recover relative cell population changes (infer cell status e.g., birth, proliferation, and, death) across timepoints, and, iii) learn cell drift and identify drift genes.

VAE projects these cells to a d-dimensional ($d \ll g$) latent space $Z_t \in \mathbb{R}^{n_t \times d}$. Then, the VAE and uDSB are jointly trained to model evolution of cellular trajectories from $t = 0$ to $t = T$. The uDSB training proceeds iteratively, alternating between forward and backward passes, with the VAE training integrated into the forward pass. In the forward pass, an optimal forward drift Q_θ is approximated by simulating a backward-SDE $\{\overleftarrow{Z}_t\}_{t=\{T,T-1,..,1\}}$, where $\overleftarrow{Z}_t \in \mathcal{R}^{n_t \times d}$ and minimizing the loss $L_{div,\theta}$ (see 3.3). The VAE is jointly trained in this forward pass by minimizing latent and reconstruction losses L_{joint} (see 3.3). Then, in the backward pass, an optimal backward drift $\widehat{Q}_{\widehat{\theta}}$ is approximated by simulating a forward-SDE $\{\overrightarrow{Z}_t\}_{t=\{T,T-1,..,1\}}$, where $\overrightarrow{Z}_t \in \mathcal{R}^{n_t \times d}$ and minimizing the loss $L_{div,\widehat{\theta}}$ (see 3.3). Additionally, to predict cell population changes, another neural network K_ω is learned which determines how cell statuses (e.g., live, dead) (see 3.3) change. Once trained, ARTEMIS learns the optimal parameters for the VAE (φ, ϕ) and uDSB $(\theta, \widehat{\theta}, \omega)$. The pseudo-code to train ARTEMIS is summarized in Algorithm S1.

Table 3.1: Three scRNA-seq datasets using for benchmarking in this paper.

Dataset	**#cells**	**#timepoints**	**Source**
Pancreatic β-cell differentiation (Pancreatic)	51,274	8	GEO (GSE114412) [152]
Zebrafish embryogenesis (Zebrafish)	38,731	12	Broad Single-Cell Portal SCP162 [46]
Epithelial-to-mesenchymal transition (EMT)	3,133	5	GEO (GSE147405) [31]

Table 3.2: Wasserstein distance between the predicted and held-out timepoints. Numbers in **bold** indicate best performance.

Datasets	**Pancreatic**		**Zebrafish**			**EMT**
	t=3	t=6	t=4	t=6	t=8	t=2
ARTEMIS	9.84±0.008	**9.05±0.009**	30.38±0.04	**28.96±0.073**	**1.3±0.034**	**5.1±0.01**
uDSB	11.209±0.011	10.81±0.009	39.98±0.18	42.21±0.15	43.23±0.05	47.26±0.12
scNODE	**9.73±0.0003**	9.22±0.0005	**29.8±0.002**	29.2±0.09	31.8±0.003	46.2±0.005
MIOFlow	10.46±0.012	10.47±0.016	31.2±0.03	31.19±0.04	34.42±0.04	45.4±0.006
Prescient	10.36±0.011	11.12±0.032	50±0.15	47.67±0.14	49.2±0.2	47.01±0.02
Prescient (w. growth rates)	10.55±0.03	12.16±0.07	-	-	-	48.83±0.01

Learning interpretable latent space with variational autoencoders

The Variational Autoencoder (VAE) [87] has been one of the most popular generative models. The basic idea of VAE can be summarized as follows: (1) VAE encodes the input data samples into a latent variable as its distribution of representation via a probabilistic encoder, which is parameterized by a neural network. (2) It then adopts the decoder to reconstruct the original input data based on the samples from the latent variable. Here, we pre-trained a VAE on scRNA-seq data for observed timepoints, where an encoder module $q(\cdot;\varphi) : \mathbb{R}^{g+1} \to \mathbb{R}^d$ maps cells concatenated with sinusoidal encoded time $(X_t||t)$ to a latent space parameterized by a Normal distribution $\mathcal{N}(\mu, \sigma^2)$, and a decoder $p(\cdot;\phi) : \mathbb{R}^{d+1} \to \mathbb{R}^g$ maps $(Z_t||t)$ back to gene expression space:

$$q_\varphi(x_t||t) = (\mu_t, \sigma_t) \tag{3.1a}$$

$$z_t \sim \mathcal{N}(\mu_t, \sigma_t^2) \tag{3.1b}$$

$$\widehat{x}_t = p_\phi(z_t||t) \tag{3.1c}$$

The encoder and decoder networks are parameterized by φ and ϕ, respectively. Then, the loss function L_{vae} minimized includes i) mean squared error (MSE) between the input and reconstructed scRNA-seq data and, ii) the Kulback-Leibler (KL) divergence between the encoder output and a standard normal prior:

$$L_{vae} = MSE(X_t, \widehat{X}_t) + \beta KL(Z_t, \mathcal{N}(0, 1)), \tag{3.2}$$

where β is a scaling factor for the KL-divergence term to ensure that the model learns robust and interpretable latent representations [69].

Modeling latent time-series with Schrödinger bridges

The Schrödinger bridge (SB) is the solution to an entropy-regularized optimal transport problem of finding the most likely evolution between two probability distributions. It seeks to find an optimal pair of forward-backward stochastic processes (SDEs) of the forms:

$$\mathrm{d}\overrightarrow{Z}_t = [f + \varepsilon^2 \nabla \log \Psi(\overrightarrow{Z}_t, t)]dt + \varepsilon \mathrm{d}W_t, \quad \overrightarrow{Z}_0 \backsim \rho_0 \tag{3.3a}$$

$$\mathrm{d}\overleftarrow{Z}_t = [f - \varepsilon^2 \nabla \log \widehat{\Psi}(\overleftarrow{Z}_t, t)]dt + \varepsilon \mathrm{d}\overline{W}_t, \quad \overleftarrow{Z}_0 \backsim \rho_T \tag{3.3b}$$

where (ρ_0,ρ_T) are the boundary distributions such that $\rho_0 = \mathcal{N}(\mu_\varphi(X_0||0), \sigma^2_\varphi(X_0||0))$, $\rho_T = \mathcal{N}(\mu_\varphi(X_T||T), \sigma^2_\varphi(X_T||T))$; $\nabla \log \Psi(Z_t, t)$ and $\nabla \log \widehat{\Psi}(Z_t, t)$ are the optimal forward and backward drifts, $\{W_t,\ \overline{W}_t \in \mathbb{R}^d\}$ are standard Wiener processes and its time reversal, and f and $\varepsilon > 0$ are base drift and diffusion coefficient, respectively. Both f and ε are constants known in prior. Moreover, the two SDEs in (3.3a), (3.3b) can be thought of as *reversed* to each other. Now, suppose Ψ,$\widehat{\Psi} \in C^{2,1}(\mathbb{R}^d, [0, T])$ solve the following coupled PDEs,

$$\begin{cases} \dfrac{\partial \Psi(z,t)}{\partial t} = -\nabla \Psi^T f - \frac{1}{2}\varepsilon^2 \Delta \Psi, \\ \dfrac{\partial \widehat{\Psi}(z,t)}{\partial t} = -\nabla \cdot (\widehat{\Psi} f) + \frac{1}{2}\varepsilon^2 \Delta \widehat{\Psi} \end{cases} \quad \text{s.t.} \quad \begin{cases} \Psi(\cdot, 0)\, \widehat{\Psi}(\cdot, 0) = \rho_0, \\ \Psi(\cdot, T)\, \widehat{\Psi}(\cdot, T) = \rho_T \end{cases} \tag{3.4}$$

where $\nabla\cdot$ is the divergence operator and Δ is the laplace operator. Then, according to the SB theory, the solution to (3.4) can be expressed through the two coupled SDEs in (3.3)[94]. Here, the optimal forward ($\nabla \log \Psi(Z_t, t)$) and backward ($\nabla \log \widehat{\Psi}(Z_t, t)$) drifts are generally unknown and can be learned through neural networks parameterized by θ,$\widehat{\theta}$, i.e.

$$Q(\cdot, \cdot; \theta) \approx \varepsilon \nabla \log \Psi(\cdot, \cdot) \qquad \widehat{Q}(\cdot, \cdot; \widehat{\theta}) \approx \varepsilon \nabla \log \widehat{\Psi}(\cdot, \cdot) \tag{3.5}$$

and is called the diffusion Schrödinger bridge (DSB). However, due to coupling constraints at the boundaries, solving (3.4) is a daunting task. Recently, [27] introduced a likelihood training framework grounded on forward-backward SDE (FB-SDE) theory, which allows to construction of likelihood objectives for training DSBs. Then the negative likelihood loss functions to solve for θ and $\widehat{\theta}$ are:

$$\begin{aligned} L_{div,\widehat{\theta}}(\overrightarrow{z}_0; \widehat{\theta}) = \int_0^T \mathbb{E}_{Z_t \sim (3.3a)}[\frac{1}{2}\|\widehat{Q}_{\widehat{\theta},t}\|^2 + \varepsilon \nabla_x \cdot \widehat{Q}_{\widehat{\theta},t} \\ + \langle \widehat{Q}_{\widehat{\theta},t}, Q_{\theta,t} \rangle \mathrm{d}t], \end{aligned} \tag{3.6a}$$

$$\begin{aligned} L_{div,\theta}(\overleftarrow{z}_0; \theta) = \int_0^T \mathbb{E}_{Z_t \sim (3.3b)}[\frac{1}{2}\|Q_{\theta,t}\|^2 + \varepsilon \nabla_x \cdot Q_{\theta,t} \\ + \langle Q_{\theta,t}, \widehat{Q}_{\widehat{\theta},t} \rangle \mathrm{d}t], \end{aligned} \tag{3.6b}$$

where $\nabla_x \cdot$ denotes the divergence operator with respect to the variable x: For any $v : \mathbb{R}^d \to \mathbb{R}^d$, $\nabla_x \cdot v(x) := \sum_{i=1}^d \frac{\partial}{\partial x_i} v_i(x)$. While DSBs assume that extreme/boundary marginals are normalized distributions, unbalanced DSBs (uDSBs) [117] relax this constraint by considering marginals with arbitrary mass i.e. by incorporating cell birth and death mechanisms. This is done by extending the state space $\mathbb{R}^d$ to its *one-point compactification*, denoted as $\widehat{\mathbb{R}}^d = \mathbb{R}^d \cup \{\infty\}$, in which the added point ∞ serves as a "cemetery" or "coffin" state. This allows for jumps in processes, where new cells are born when state changes from $\infty \to \mathbb{R}^d$ or existing cells are killed when $\mathbb{R}^d \to \infty$. In this paper, we assume that in the absence of prior information about cell population changes, the number of cells (n_t) collected at each timepoint represents the relative cell population. Also, we assume that any variability introduced by sequencing technologies, such as incomplete sequencing of cells, is negligible when modeling the relative cell population changes over time. Then, in addition to estimating θ and $\widehat{\theta}$, a posterior kill rate is learned using another neural network $K(\cdot;\omega)$:

$$k` \approx k(\overrightarrow{z}, t) K_\omega(t), \tag{3.7}$$

where $k(z,t) > 0$ is the prior kill rate. We follow uDSB to define the prior kill rate for a cell, which is defined as the ratio of the number of features deviating by more than two standard deviations (from the mean of expression of cells from the next measured timepoint) to the total number of features for that cell. If fewer than 20% of the features deviate, the prior kill rate is set to 0. The loss function to optimize ω is given by:

$$\begin{aligned} L_\omega = \mathbb{E}_{(X_t, A_t)}[\sum_{i \in I_t} | \int_0^i ((1 - A_t) b K_\omega(t) \\ - A_t K_\omega(t)) \mathrm{d}t - \frac{n_i - n_0}{\max_i n_i} |], \end{aligned} \tag{3.8}$$

where $A_t = [a_{t,1}, ..., a_{t,n_t}] \in \mathcal{R}^{n_t}$ where $a_{t,i} \in \{0, 1\}$ for all $i = 1, ..., n_t$ such that $a_{t,i} = 1$ indicates live and $a_{t,i} = 0$ indicates dead cell status for some cell i, b is the birth rate (i.e. negative death rate) and I_t is the set of intermediate timepoints. This loss function helps the network track cell population changes by estimating cell births and deaths at each time step and adjusting predictions to match observed changes. The first two terms enforce that the mass change predicted by uDSBs in each time interval [0,i] matches the empirical change $(n_i - n_0)$ for observed timepoints $t \in I_t$. Additionally, we use the Euler-Maruyama discretization [117, 27] to approximate the SDEs from (3.3):

$$\Delta \overrightarrow{Z}_t = [f + \varepsilon Q_\theta(\overrightarrow{Z}_t, t)]\Delta t + \varepsilon \Delta W_t, \quad \overrightarrow{Z}_0 \backsim \rho_0 \tag{3.9a}$$

$$\Delta \overleftarrow{Z}_t = [f - \varepsilon \widehat{Q}_{\widehat{\theta}}(\overleftarrow{Z}_t, t)]\Delta t + \varepsilon \Delta \overleftarrow{W}_t, \quad \overleftarrow{Z}_0 \backsim \rho_T \tag{3.9b}$$

where the interval $[0, T]$ is discretized into 100 steps. Then the descritized loss $L_k(\omega)$ is:

$$\begin{aligned} L_\omega = \mathbb{E}_{(\overrightarrow{Z}_t, A_t)}[&\sum_{i\in I_t}\sum_{t=0}^{i}(1-A_t)\lceil b_{\leftarrow\infty}(\overrightarrow{Z}_t)\rfloor - A_t\lceil k_{\rightarrow\infty}(\overrightarrow{Z}_t)\rfloor) \\ &- \frac{n_t - n_0}{\max_i n_i} + \sum_{t=0}^{i}(b_{\leftarrow\infty}(\overrightarrow{Z}_t) - \lceil b_{\leftarrow\infty}(\overrightarrow{Z}_t)\rfloor) \\ &+ \sum_{t=0}^{i}(k_{\rightarrow\infty}(\overrightarrow{Z}_t) - \lceil k_{\rightarrow\infty}(\overrightarrow{Z}_t)\rfloor)], \end{aligned} \tag{3.10}$$

where $\lceil x \rfloor$ denotes that x is clipped to a unit interval. The last two regularization terms are added to penalize transition probabilities $k_{\rightarrow}\infty$ and $q_{\leftarrow}\infty$ greater than 1, where

$$\begin{aligned} k_{\rightarrow\infty}(\overrightarrow{z}_t) &= \mathbb{P}(\overrightarrow{Z}_{t+1} = \infty | \overrightarrow{Z}_t = \overrightarrow{z}_t) = k(\overrightarrow{z}_t, t) K_{\omega,t}(\overrightarrow{z}_t)\Delta t \\ b_{\leftarrow\infty}(\overrightarrow{z}_t) &= \mathbb{P}(\overrightarrow{Z}_{t+1} = \overrightarrow{z}_t | \overrightarrow{Z}_t = \infty) = b(\overrightarrow{z}_t, t) K_{\omega,t}(\overrightarrow{z}_{t+1})\Delta t \end{aligned} \tag{3.11}$$

The discretized divergence losses (3.6a), (3.6b) become:

$$\begin{aligned} L_{div,\widehat{\theta}}(\overrightarrow{z}_0; \widehat{\theta}) = \Delta t \sum_{i=0}^{T} A_i (&\frac{1}{2}||\widehat{Q}_{\widehat{\theta},i}(\overrightarrow{Z}_i)||^2 + \varepsilon \nabla_x \cdot \widehat{Q}_{\widehat{\theta},i}(\overrightarrow{Z}_i) \\ &+ \langle \widehat{Q}_{\widehat{\theta},i}(\overrightarrow{Z}_i), Q_{\theta,i}(\overrightarrow{Z}_i)\rangle), \end{aligned} \tag{3.12a}$$

$$\begin{aligned} L_{div,\theta}(\overleftarrow{z}_0; \theta) = \Delta t \sum_{i=0}^{T} A_i (&\frac{1}{2}||Q_{\theta,i}(\overleftarrow{Z}_i)||^2 + \varepsilon \nabla_x \cdot Q_{\theta,i}(\overleftarrow{Z}_i) \\ &+ \langle Q_{\theta,i}(\overleftarrow{Z}_i) \widehat{Q}_{\widehat{\theta},i}(\overleftarrow{Z}_i)\rangle), \end{aligned} \tag{3.12b}$$

where $\overrightarrow{Z}_i \sim$ (3.9a) and $\overleftarrow{Z}_i \sim$ (3.9b). The sampling procedure for forward and backward SDEs with birth and death mechanisms is summarized in Algorithm S2 where the predicted kill rate r‘ determines if cells are born, proliferate or die.

To learn optimal forward and backward drifts $(\theta^*, \widehat{\theta}^*)$, the iterative proportional fitting (IPF) algorithm [48, 129], which is the dynamic version of the sinkhorn algorithm, [33] has been used widely. It works by alternating between two steps, (i) a forward pass that adjusts the predicted distribution of a forward SDE ((3.9a)) to match the terminal distribution i.e. $\overleftarrow{Z}_0 \sim \rho_T$, and (ii) a backward pass that adjusts the predicted distribution of a backward SDE ((3.9b)) to match the initial distribution i.e. $\overrightarrow{Z}_0 \sim \rho_0$. This process is repeated iteratively, with each iteration bringing the predicted distributions closer to satisfying marginal distributions. The training details are summarized in Appendix A.2(Supplementary Note S1 and Algorithm S1). For more details about uDSBs, we refer readers to [117].

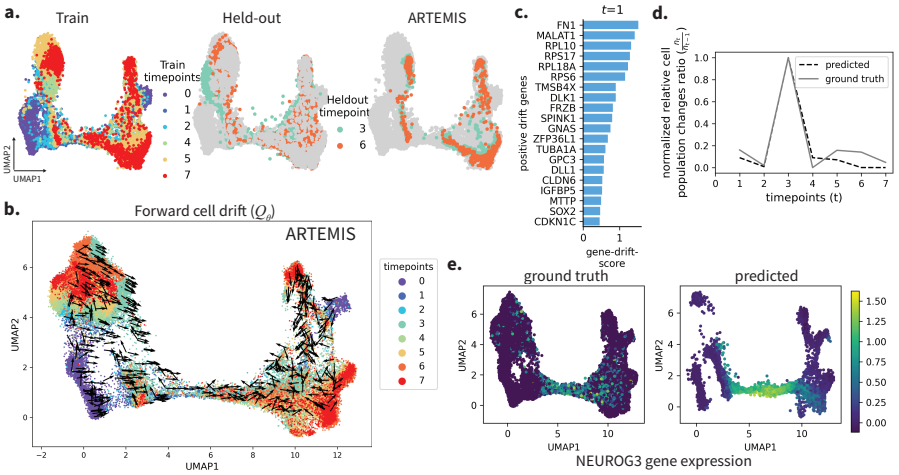

Figure 3.2: **Application to pancreatic β-cell differentiation spanning eight days (0-7).** a.) 2D UMAP to show ARTEMIS's performance on held-out timepoints (3,6). b.) Visualization of the drift inferred by ARTEMIS trained on six timepoints. c.) Top 20 drift-genes identified for $t = 1$ from the forward drift Q^θ. d.) Comparison of normalized ratios of relative cell population changes between ground truth and ARTEMIS-predicted cell statuses. e.) Ground truth vs. predicted gene expression of transient TF *NEUROG3*.

Joint optimization of VAE and Schrödinger bridges for smooth time-series dynamics in latent space

Following pre-training of the VAE on gene expression data from measured timepoints, we jointly train the VAE and the uDSB to learn a smooth latent space for interpolation to unmeasured intermediate timepoints. As detailed in 3.3, the uDSB is trained iteratively through forward and backward passes to approximate optimal drifts while adhering to marginal distributions. However, this training process overlooks cellular profiles at intermediate timepoints. To address this, we introduce a latent loss term into the SB training, penalizing discrepancies between gene expression predictions in the latent space generated by the VAE and the uDSB. Additionally, the VAE is jointly trained with the forward pass, incorporating two additional loss functions into $L_{div,\theta}(z_T;\theta)$ to minimize the distance: i) between VAE-encoded latents $Z_{\varphi,t}$ and uDSB-predicted latents Z_t, and ii) between the ground truth and VAE-reconstructed gene expression:

$$L_{joint} = W_2(Z_{\varphi,t}, \overrightarrow{Z}_t) + W_2(X_t, p_\phi(Z_{\varphi,t})), \tag{3.13}$$

where $W_2(\mu,\nu)$ calculates the 2-Wasserstein distance between empirical distributions μ and ν.

Model Outputs

To reconstruct a complete trajectory from $t = 0$ to $t = T$, X_0 is input to the trained model, which is mapped to the latent space by the encoder q_φ. Then a forward trajectory $\overrightarrow{Z}_t$ is sampled using (3.9a) up to $t = T$. Then $\overrightarrow{Z}_t$ is mapped back

to the gene expression using the decoder p_ϕ computing expression for measured(t) and unmeasured($\tilde{t}$) timepoints: $\widehat{X}_t, \widehat{X}_{\tilde{t}}$. In addition to expression, ARTEMIS also outputs the cell status $A_t, A_{\tilde{t}}$ for cells in the reconstructed trajectory.

Datasets and preprocessing

We benchmarked ARTEMIS on three time-series scRNA-seq datasets (Table 3.1). The pancreatic dataset spans Days 0 to 7, capturing stage 5 human in vitro pancreatic β-cell differentiation [152]. The zebrafish dataset covers embryogenesis across twelve stages, measured in hours post-fertilization (hpf) [46]. Lastly, the EMT dataset involves an A549 lung cancer cell line treated with TGFB1 to induce epithelial-to-mesenchymal transition (EMT), sampled at five timepoints from 8 hours to 1 week post-treatment [31]. For all datasets, counts were normalized using depth scaling, ensuring that the total counts for each cell across all genes were consistent. This was followed by log1p normalization (log-transformation after adding one) to stabilize variance. These steps prevented data leakage between training and held-out data. Next, we identified 2,000 highly variable genes (HVGs) from the training data and applied this selection to the held-out samples. Preprocessing was performed using the *Scanpy* package [160]. For the EMT dataset, we additionally applied z-score normalization using the *scikit-learn* package [118], as we perform perturbation analysis on this data. To simplify computations, we relabeled timepoints as consecutive integers starting from 0. For the pancreatic dataset, we filtered out genes correlated with *TOP2A* ($r > 0.15$) following Yeo et al. (2021) before training resulting in 1922 HVG genes.

3.4 Results

Human in vitro β-cell differentiation in pancreas

ARTEMIS was first applied to a pancreatic β-cell differentiation dataset with eight timepoints [152]. To evaluate ARTEMIS's predictive performance, we withheld two timepoints together–3 and 6–during training for comparison across methods, as 3 is central to differentiation and 6 is near the terminal stage. ARTEMIS demonstrated superior accuracy in reconstructing both timepoints, with better generalization to the later timepoint ($t = 6$), as measured by average Wasserstein distance (Table 3.2). This is illustrated by a 2D UMAP visualization of the predicted gene expressions (Figure 3.2a., Figure A.14a.).

To further analyze the temporal evolution of cellular trajectories, we visualized the forward drift ($Q_\theta(\cdot, \cdot)$) inferred by ARTEMIS, representing deterministic trends in gene expression dynamics. ARTEMIS accurately modeled the progression of cells from progenitor states toward terminal states, such as exocrine and neurog3_early populations (t=0, lower left in Figure 3.2b.), by aligning drift directions with biological expectations. We compare this drift to PRESCIENT, as it also infers cellular

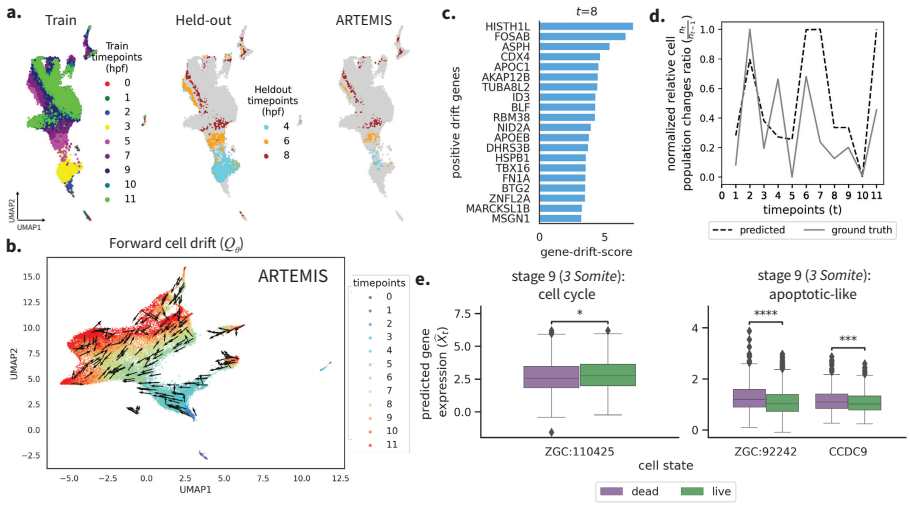

Figure 3.3: **Application to zebrafish embryogenesis data across twelve stages (i.e. hours post fertilization (hpf)).** a.) 2D UMAP to show ARTEMIS's performance on held-out timepoints (4,6,8). b.) Visualization of the drift inferred by ARTEMIS trained on nine timepoints. c.) Top 20 drift-genes identified for $t = 8$. d.) Comparison of normalized ratios of relative cell population changes between ground truth and ARTEMIS-predicted cell statuses as live. e.)Boxplots showing gene expression of DE genes between cells predicted as live and dead by ARTEMIS during the interval $t = 9$ to $t = 10$.

drift. However, PRESCIENT's inferred drift often misaligns directions, particularly in regions of bifurcation (Figure A.14b.).

To identify the key genes driving these cellular transitions, we performed drift-gene analysis by projecting the latent forward drift ($Q^\theta(\cdot,\cdot)$) onto the gene expression space. Assuming genes with positive drift scores as drivers of cellular trajectories, we identified the top 20 drift-genes at each timepoint (Figure 3.2c., Figure A.14d., see Appendix A.2(Supplementary Notes S5)). At $t = 1$, genes such as *MALAT1* ($p < 2.9e^{-23}$), *FN1* ($p < 1.7e^{-34}$), and *SOX2* ($p < 2.9e^{-2}$) were among the top-ranked and also previously known markers of stage 5 progenitor cells ([152]).

We then evaluated ARTEMIS's ability to recover cellular population changes between $t = 0$ and $t = 7$. Relative cell population change ratios for groundtruth were calculated as $(\frac{n_t}{n_{t-1}})$ and normalized. Similarly, for ARTEMIS, these were given by $(\frac{\sum A_t}{\sum A_{t-1}})$ and normalized. We see that the trends captured by ARTEMIS closely matched groundtruth, suggesting that ARTEMIS can recover relative cell population changes for unmeasured timepoints (Figure 3.2d.) and inferred cell statuses capturing the increase and decrease in relative cell populations as birth and proliferation, and death, respectively (Figure A.15c.).

Additionally, ARTEMIS predicted the transient expression of *NEUROG3*, a critical transcription factor for endocrine induction and β-cell differentiation (Figure 3.2d.) [152] accurately. These results demonstrate ARTEMIS's ability to integrate gene expression and population dynamics while identifying key genes driving cellular transitions.

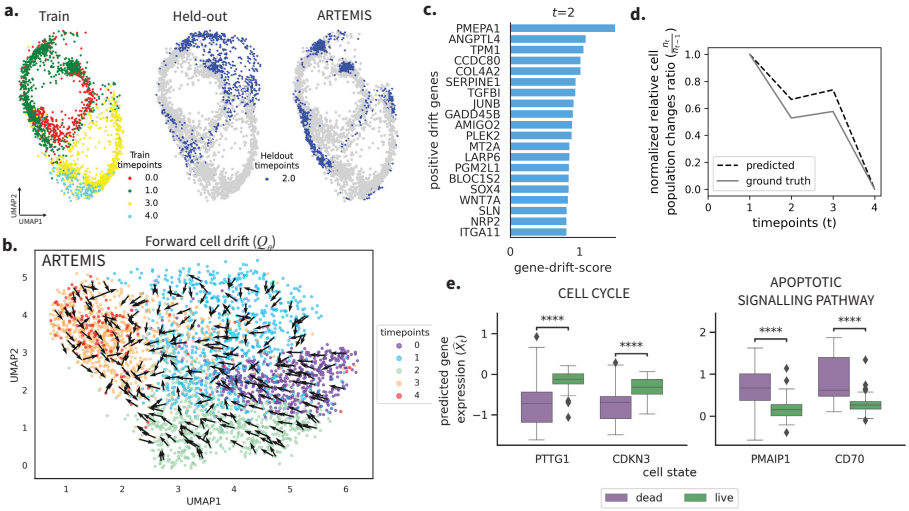

Figure 3.4: **Application to A549 lung cancer cells undergoing TGFB1-induced EMT spanning five timepoints.** a.) 2D UMAP to show ARTEMIS's performance on held-out timepoint (4). b.) Visualization of the drift inferred by ARTEMIS trained on four timepoints. c.) Top 20 drift-genes identified for $t = 2$. d.) Comparison of normalized ratios of relative cell population changes between ground truth and ARTEMIS-predicted cell statuses as live. e.) Boxplots showing gene expression of DE genes between cells predicted as live and dead by ARTEMIS during the interval $t = 3$ to $t = 4$.

Developmental stages underlying zebrafish embryogenesis

We next applied ARTEMIS to a zebrafish embryogenesis dataset spanning twelve developmental stages [46]. To evaluate prediction accuracy, three timepoints ($t = 4, 6, 8$) were held out together during training as previously benchmarked by scNODE. ARTEMIS outperformed other methods in reconstructing these timepoints, with better generalization to later stages (Table 3.2), further illustrated by 2D UMAP visualizations of the predicted gene expressions (Figure 3.3a., Figure A.16a.).

To analyze developmental trajectories, we visualized the forward drift inferred by ARTEMIS and compared to PRESCIENT. ARTEMIS successfully modeled the deterministic evolution of cells through distinct developmental phases, aligning the drift consistently toward later stages of embryogenesis (Figure 3.3b., PRSCIENT: Figure A.16b.). Drift-gene analysis highlighted key regulators of zebrafish development at various stages (Figure 3.3c., Figure A.17d.). For example, at $t = 8$, top drift genes identified included *CDX4* ($p < 2.9e^{-6}$), *APOC1* ($p < 2.6e^{-9}$), and *TBX16* ($p < 2.5e^{-17}$), known for their roles in Tailbud formation. Similarly, genes such as *ID3* ($p < 8.9e^{-25}$) and *FN1A* ($p < 2.8e^{-60}$) are known for their role in Prechordal Plate lineage, and *ASPH* ($p < 1.7e^{-28}$) and *APOEB* ($p < 7.5e^{-12}$) in the Endoderm lineage. ([46]).

ARTEMIS also recovered cellular population changes across $t = 0$ to $t = 11$, closely matching the ground truth for relative cell population changes (Figure 3.3d.) and inferred cell statuses (Figure A.17c.). We also conducted DE analysis dur-

ing the interval $t = 9$ to $t = 10$, where ARTEMIS predicted enough number of cells as dead. Here, ARTEMIS identified genes associated with the cell cycle, such as *ZGC:110425*($p < 0.05$), as enriched in cells predicted live, while apoptotic-like genes, including *ZGC:92242*($p < 1e^{-4}$) and *CCDC9*($p < 1e^{-3}$), exhibited higher expression in cells predicted dead (Figure 3.3e.)[46]. This shows that ARTEMIS provides meaningful predictions of cell statuses, linking gene expression patterns to cellular states such as proliferation and apoptosis during key developmental dynamics.

Epithelial-to-mesenchymal transition in TFGB1-induced A549 lung cancer cells

Finally, we applied ARTEMIS to an Epithelial-to-mesenchymal (EMT) dataset of A549 lung cancer cells treated with TGFB1, spanning five timepoints. To evaluate predictive performance, $t = 2$ was held out during training as it is the central timepoint in the measured EMT process. ARTEMIS achieved the lowest average Wasserstein distance, outperforming other methods in reconstructing the held-out timepoint (Table 3.2), illustrated using a 2D UMAP visualization (Figure 3.4a., Figure A.17a.).

The forward drift inferred by ARTEMIS successfully captured the progression of cells toward later timepoints, modeling the deterministic trends associated with EMT (Figure 3.4b.). In contrast, PRESCIENT struggled to align directions accurately in complex regions (Figure A.17b.). Drift-gene analysis identified key regulators of EMT at each timepoint (Figure 4c., Figure A.17d.). At $t = 2$, genes such as *COL4A2* ($p < 5.7e^{-10}$), *PMEPA1* ($p < 7.5e^{-9}$), *SERPINE1*($p < 6.2e^{-4}$), and *TPM1*($p < 1.9e^{-15}$) that were identified as top drift genes, are also known as EMT hallmark genes [98].

ARTEMIS also captured relative cell population changes across $t = 0$ to $t = 4$, closely matching the ground truth trends (Figure 3.4c.) and inferred cell statuses (Appendix A.2 (Supplemetary Note S3c.)). Differential expression (DE) analysis during $t = 3$ to $t = 4$, where most cells were predicted as dead, revealed distinct patterns. Genes previously associated with the GOBP CELL CYLCE pathway, such as *PTTG1*($p < 1e^{-4}$) and *CDKN3*($p < 1e^{-4}$), were enriched in cells predicted as liv, while genes previously known in the GOBP APOPTOTIC SIGNALLING PATHWAY, including *PMAIP1*($p < 1e^{-4}$) and *CD70*($p < 1e^{-4}$), exhibited higher expression in cells predicted as dead (Figure 3.4e.) [98].

To further validate our findings and explore the role of identified drift genes *TPM1* and AMIGO2, we introduced *in silico* perturbations on cells at timepoint $t = 2$ by introducing varying levels of overexpression (5, 10, 15, 20, 25) or underexpression ($-25, -20, -15, -10, -5$) of the drift genes (Figure 3.5, Figure A.18, see Appendix A.2 (Supplementary Note S6)). The trained ARTEMIS model was initialized by 2000 randomly sampled cells from $t = 2$, and allowed to reconstruct remaining trajectory up to terminal timepoint $t = 4$. At $t = 2$, model was initialized with unperturbed cells and cells with perturbations. This was repeated for 10

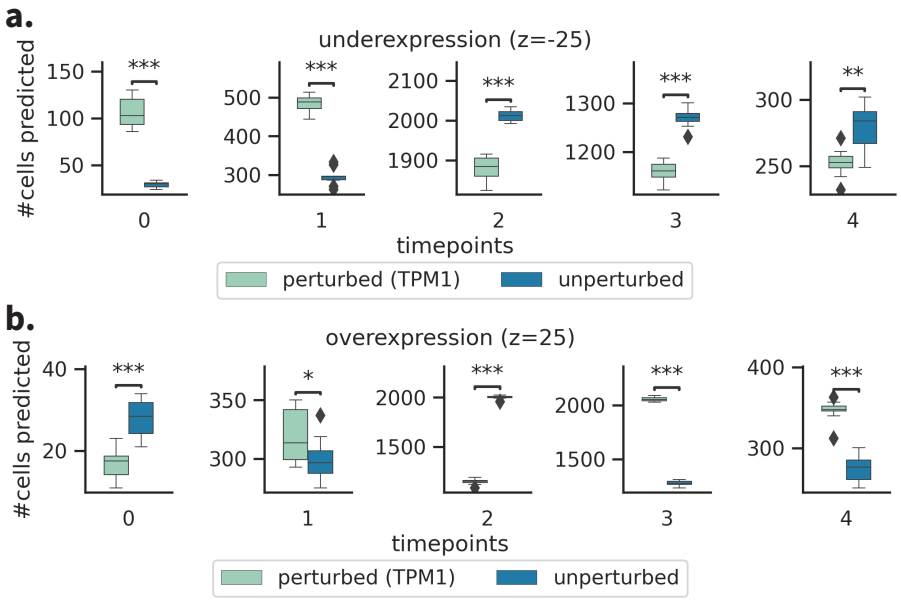

Figure 3.5: **Perturbation analysis on A549 lung cancer cells undergoing TGFB1-induced EMT.** ARTEMIS was initialized with 2000 cells sampled from $t = 2$, either unperturbed or perturbed by TPM1 expression changes, and used to simulate trajectories to the terminal timepoint. An MLP classifier assigned cells to specific timepoints. Boxplots show the number of cells assigned to each timepoint in perturbed vs. unperturbed settings. Differences were assessed using a two-sided t-test at $p < 0.05$ (see Appendix A.2 (Supplementary Note S6)). (a) Underexpression (-25 perturbation), (b) Overexpression ($+25$ perturbation).

trials. An MLP classifier, trained on ground truth latents inferred by the VAE, was used to classify the cells from the predicted trajectory into five timepoints. For each timepoint, a two-sided t-test ($p < 0.05$) compared the distribution of cells between perturbed and unperturbed groups. When underexpressed, there was an increase in cells from earlier timepoints ($t = 0, 1$) and a decrease in cells from later timepoints in perturbed trajectories, compared to the unperturbed trajectories ($p < 0.001$, Figure 3.5a). Conversely, overexpression resulted in increased cell populations at later timepoints ($t = 3, 4$) and a decrease at earlier timepoints ($p < 0.05$) in perturbed trajectories (Figure 3.5b). Similar patterns were observed across other perturbation levels (Figure A.18). These findings suggest that the drift genes identified by ARTEMIS not only reflect deterministic trends underlying EMT but can also influence cellular trajectories when perturbed.

Chapter 4

Personalized Single-Cell Transcriptomics Analysis in Alzheimer's Disease

4.1 Abstract

Understanding the heterogeneity of brain diseases requires approaches that capture population-level variation in molecular and cellular mechanisms. Functional genomics has emerged as a key strategy for investigating mechanisms such as gene regulation and cell–cell communication in Alzheimer's disease (AD). While recent large-scale single-cell studies have revealed population-level gene expression trends, modelling functional genomic variation at the individual donor level remains challenging. The PsychAD project generated single-nucleus RNA-seq data from 1,494 human brains, comprising over 6.3 million nuclei, spanning a range of clinical phenotypes and neuropsychiatric symptoms in AD. Leveraging this dataset, we performed personalized functional genomics analyses, capturing each donor's cell type interactions and gene regulatory networks. Using a knowledge-guided graph neural network, we learned latent representations of functional genomics (embeddings) and scored the importance of cell types, genes, and their interactions for each donor. In contrast to population-averaged methods, our framework preserves individual donor variation in cell types and genes and their interactions. The embeddings improved phenotype classification and enabled discovery of subtypes and disease progression trajectories in AD. Our importance scores identified significant inter-donor differences in gene regulation and prioritized personalized functional genomic features. We validated our findings in external cohorts, demonstrating robustness and generalizability. All results are available via iBrainMap, an open-source framework and personalized functional genomics atlas for AD. Our work provides a scalable approach to model population-scale functional genomic variation and offers a resource for mechanistic insights and precision targeting in brain diseases.

4.2 Introduction

The human brain exhibits higher transcriptomic complexity than most other tissues [73, 170], making it susceptible to a wide range of neurodegenerative and

neuropsychiatric diseases. These diseases are heterogeneous, with a broad spectrum of symptoms, varying severity levels, and phenotypic differences [85, 164] that stem from diverse cellular and molecular mechanisms [113, 80]. In the context of Alzheimer's disease (AD), large-scale single-cell RNA-seq datasets across individuals (e.g., ROSMAP [107] and SEA-AD[68]) have led to the study of gene expression variation at the cell type level [107, 109, 56]. These studies have expanded beyond binary case-control comparisons by incorporating detailed AD phenotypic stratification (e.g., resilience, pathology-cognition) and individual-level metrics. Complementing these studies, functional genomics has emerged to investigate gene regulatory mechanisms, linking genetic variants to changes in gene expression and downstream cellular functions underlying these diseases [170, 79, 122]. For instance, gene regulatory networks (GRNs), which can model interactions between genes (transcription factor to target genes) at a cell type level, have identified key genes and functions associated with AD phenotypes [58, 44, 158]. Similarly, several population studies have shown that cell-cell communication is an essential component in AD (e.g., via cell type interactions such as astrocyte-microglia crosstalk in β-amyloid pathology[97] and neuroinflammation[11]). However, how these interactions vary between individuals remains unclear. While many studies have pooled cells across individuals to identify regulatory mechanisms, fewer have aimed to construct individualized functional genomic models for AD. Here, we define "personalized functional genomics" as the modeling of cell type interactions and gene regulatory networks at the individual donor level, informed by each donor's single-cell transcriptomic profile and known disease-related priors.

Large amounts of individual data are essential to capture the complexities and variability of the functional genomics present across populations in AD. Recent developments in single-cell sequencing technology have provided opportunities to characterize cell diversity, reveal the genomic landscape, and understand disease heterogeneity of the human brain at single-cell resolution [108, 88, 52]. To explore neurodegenerative and neuropsychiatric disease mechanisms, the PsychAD Consortium (Supplementary Note "PsychAD dataset") has generated population-level single-nucleus RNA sequencing (snRNA-seq) data from 1,494 donors across 6.3 million nuclei from the human dorsolateral prefrontal cortex (DLPFC) brain region [92, 51]. The dataset includes donors with neurodegenerative and neuropsychiatric disorders such as AD, schizophrenia (SCZ), and diffuse Lewy body disease (DLBD), either as standalone diagnoses or in combinations of multiple diagnoses. For a subset of donors, detailed information on neuropsychiatric symptoms (NPSs), such as depression, is available. Additionally, quantitative measurements of AD-related phenotypes, including BRAAK stages (neurofibrillary tangles) and cognitive impairment, are available for all donors diagnosed with AD and the majority of neurotypical controls as well.

Analyzing such large-scale population data typically requires emerging computational approaches to discover personalized functional genomic information that traditional methods cannot capture. Several personalized analyses in brain diseases

have used brain imaging and gene co-expression networks to capture individual-level variation [84, 146, 95, 131, 102, 150, 104]. For example, a recent approach called Dozer[104] identified individual variations in immune response among AD using personalized gene co-expression network analysis. Another study identified personalized network patterns and driver genes in various cancers by building individual-specific networks from gene expression data[102]. However, these approaches rely on correlation-based measures to infer co-expression patterns that overlook gene regulatory interactions and also lack the resolution to capture cell-type-specific interactions driving individual variation in AD. While traditional machine learning methods have provided valuable insights in identifying key genes and biomarkers and reconstructing cellular trajectories underlying AD and neurodegeneration using single-cell data [88, 53, 2], recent graph learning-based approaches have shown promise in modeling these complex interactions. For instance, approaches like SIMBA[25], scGNN[157], and GLUE[20] learn cell embeddings which are then used for cell clustering [25, 157], gene regulation inference[20, 105], and multimodal integration [157, 20, 105]. These approaches mainly focus on pooling cells from the population to learn embeddings and do not preserve intra-individual functional genomic interactions, such as cell-cell communication and gene regulatory networks.

To address this, we investigated population-scale single-cell transcriptomic data, primarily from the PsychAD cohort, and developed a graph learning framework to analyze personalized functional genomics. This approach advances the biological interpretability of prior network-based models mentioned above by capturing donor-level functional genomic patterns and prioritizing cell types, genes, and their interactions for AD and clinical phenotypes. We demonstrated the robustness of our results using independent cohorts. We also identified genetic variants associated with cell-type-level gene regulation for 27 emerging brain cell subclasses[92]. We provide a computational framework for general use, and our results are consolidated into a personalized functional genomic atlas accessible online using a webapp.

4.3 Results

Personalized functional genomic atlas across brain disease phenotypes

We present iBrainMap, a computational framework for analyzing and capturing functional genomic variations across donors by modeling cell type interactions and gene regulatory networks from population-scale snRNA-seq data (e.g., PsychAD). This offers a refined alternative to pooled or group-averaged network approaches for phenotype classification, population subtyping, and prioritization of functional genomic information at the donor level (Figure 4.1a). The iBrainMap framework first constructs a personalized functional genomic graph (PFG) for each donor, integrating cell type interactions and gene regulatory networks. Each PFG is a directed graph, with nodes representing cell types, transcription factor genes (TFs), and target genes (TGs), and edges denoting cell type interactions and cell type regulatory links (TF to TG). We then enhance each PFG by incorporating prior biological

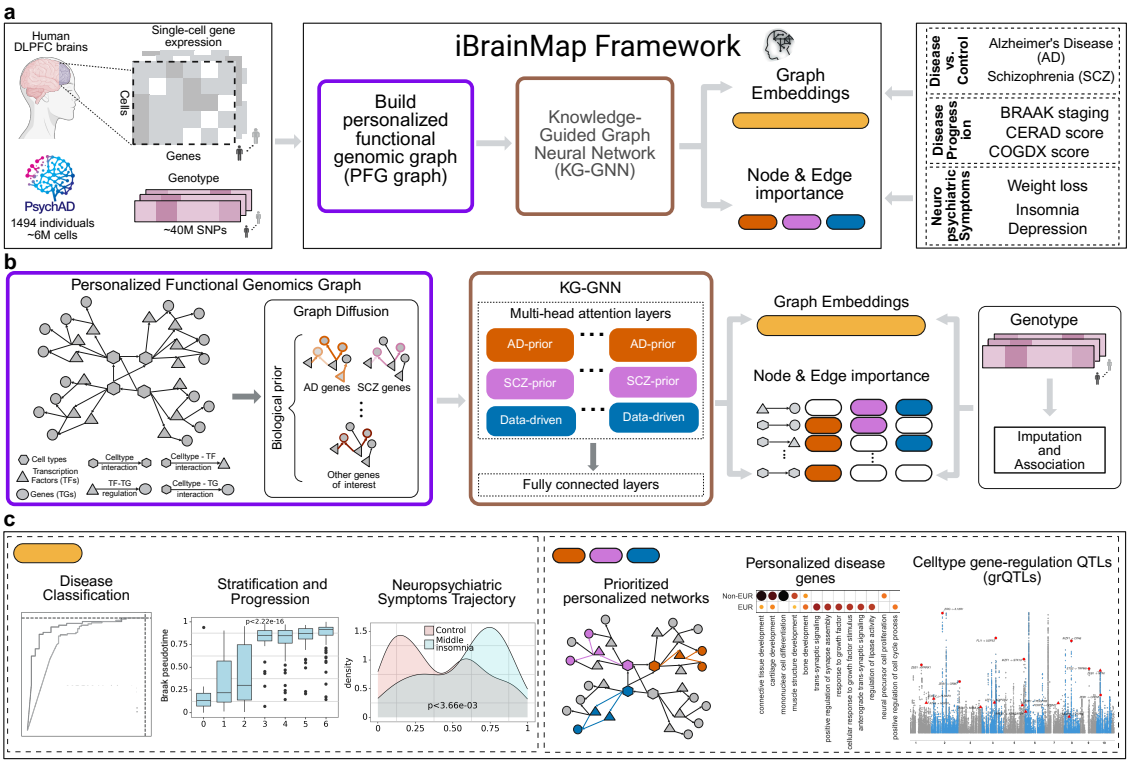

Figure 4.1: **Personalized functional genomic atlas and analysis for brain disease phenotypes.** a, The iBrainMap framework uses the PsychAD21 snRNA-seq data of 1,494 human brain donors to construct personalized functional genomic graphs (PFGs), enabling donor-level analysis of cell type interactions and gene regulation associated with brain diseases, including Alzheimer's Disease (AD) and Schizophrenia (SCZ). b, The iBrainMap framework first constructs a PFG and then applies graph diffusion to integrate known disease genes to generate bio-diffused PFGs for each donor. The KG-GNN, a multi-head graph attention model, uses the bio-diffused PFGs of each donor to learn graph embeddings for AD vs. control classification, while also prioritizing nodes and edges for AD (through learned AD prior, SCZ prior, and data-driven). Genotype data are used to associate genetic variants with gene regulation and to impute donor graph embeddings. These imputed embeddings can then be used in cohorts that only have genetic information for phenotype predictions. More details in Methods, Supplementary Notes 1-2. c, Using graph embeddings, we can classify donors across disease phenotypes and stratify them into potential novel subtypes. Using prioritized edges, we can prioritize phenotype-associated personalized networks, identify known and novel disease genes, and perform cell-type-level gene-regulation QTLs (grQTLs) that link associated SNPs with gene regulatory links.

knowledge of known disease genes (e.g., AD and SCZ genes[121, 155]) to model potential disease-relevant interactions, resulting in bio-diffused PFGs where edges associated with disease gene nodes are assigned higher weights (Supplementary Note 1, Figure A.24). These bio-diffused PFGs are then used as input to our knowledge-guided graph neural network (KG-GNN) to classify AD vs. control and generate two key outputs for each donor: latent representations of functional genomics (graph embeddings) and importance scores for PFG nodes and edges (Figure 4.1b, Supplementary Note 2, Figure A.25).

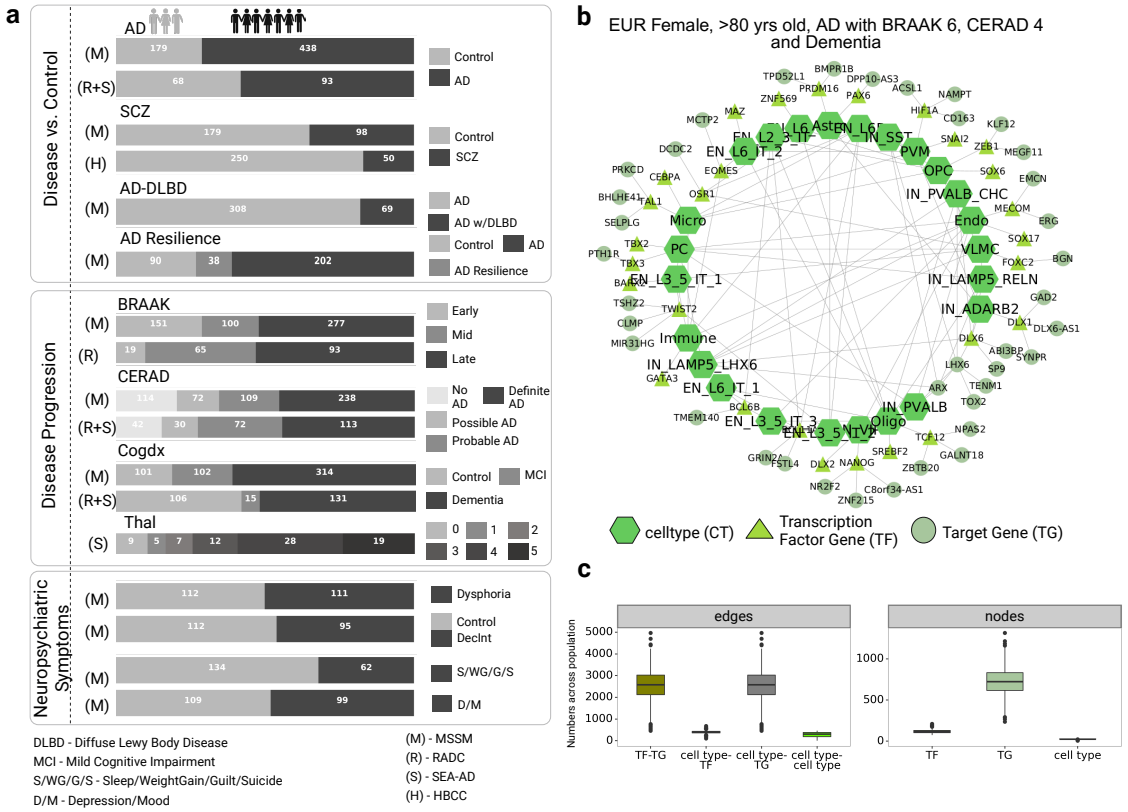

Figure 4.2: **Multi-cohort snRNA-seq data for personalized functional genomic analyses.** a, Donors from multi-cohort AD studies: PsychAD (Mount Sinai Brain Bank/MSSM (M) (n=1,042), Rush Alzheimer's Disease Center/RADC (R) (n=152), Seattle Alzheimer's Disease Brain Cell Atlas/SEA-AD (S) (n=80), and NIMH-IRP Human Brain Collection Core/HBCC (H) (n=300). Data summary of phenotypic donors used in this study divided into three levels: disease vs. control (AD vs. control, SCZ vs. control, AD-DLBD vs. control, AD-resilient vs. AD-strict vs. Control), Disease Progression (BRAAK stages, CERAD, Cogdx), and Neuropsychiatric symptoms (Dysphoria, DecInt (Anhedonia), Sleep/Weight Gain/Guilt/Suicide (S/WG/G/S), Depression/Mood (D/M)); each horizontal bar is a phenotype contrast. b, Part of a personalized functional genomics graph (PFG) for a donor from the Mount Sinai Brain Bank/MSSM (M) cohort. The donor is female with European ancestry, age >80 years old, diagnosed with AD, showing pathology for BRAAK stage 6 and CERAD score 4 (definite AD), and diagnosed with Dementia (Supplementary Note 3). Nodes can be cell types, transcription factors (TFs), and target genes (TGs); edges represent cell type interactions and cell-type-specific regulatory links (TF to TG). c, Number of PFG edges (left) and nodes (right), across 1,494 donors, constructed by iBrainMap.

Using graph embeddings, we can classify donors according to disease-related phenotypes, including case-control status, pathology capturing different stages of AD, and presence of NPSs. Additionally, we can perform population subtyping to identify novel subgroups capturing different disease stages. With pseudotime analyses of the graph embeddings, we can uncover several population trajectories associated with AD pathology, cognitive status, and NPSs (Figure 4.1c, left). The importance scores can help prioritize disease-associated functional genomics for each

donor. These prioritized nodes and edges can then identify significant subnetworks and uncover potential biomarkers associated with AD phenotypes, including cell types and regulatory elements (TFs, TGs). Additionally, the genetic variants can be associated with cell type gene regulatory network changes across donors, i.e., gene regulatory QTLs (grQTLs), providing novel functional genomic insights linking SNPs to TF-TG regulatory mechanisms as opposed to linking SNPs to genes in traditional QTL analysis (Figure 4.1c, right).

Finally, we used the PsychAD data to build the framework and independently validated our results using external datasets, ROSMAP[107] and SEA-AD[68]. The PsychAD donors were obtained from Mount Sinai NIH Brain Bank and Tissue Repository (MSSM, n=1,042 donors), NIMH-IRP Human Brain Collection Core (HBCC, n=300 donors), and Rush Alzheimer's Disease Center (RADC, n=152 donors). The phenotypes of these donors were categorized into three levels: disease vs. control, AD progression, and NPSs (Supplementary Note 3, Figure 4.2a). An illustrative example of a personalized functional graph (PFG) is shown in Figure 4.2b, and the distribution of the PFG node and edge counts across all donors in PsychAD are shown in Figure 4.2c.

Phenotype classification and subtyping using personalized functional genomics

We first trained the KG-GNN model for AD vs. control classification using donor PFGs by learning their graph embeddings. Here, we used five-fold cross-validation to identify the optimal hyperparameter configuration (Figure A.26a-b,Tabl A.34). We trained the final model using the optimal hyperparameters on a subset of donors from the AD (n=438) vs. control (n=179) contrast in MSSM cohort (Supplementary Note 4). Our KG-GNN model achieved high classification performance, with an area-under-the-curve (AUC) of 0.913 on the held-out donors from MSSM (AD: n=62 vs. control: n=30), and a combined AUC of 0.808 on independent donors from RADC and SEA-AD cohorts (AD: n=93 vs. control: n=68) (Figure 4.3a, Figure A.27a).

We benchmarked our KG-GNN model against traditional machine learning models, state-of-the-art graph learning models (GAT and GCN), and two recent graph embedding approaches (SIMBA[25] and scGNN[157], Supplementary Note 5). We outperformed traditional machine learning models which were trained on average cell type gene expression (Figure A.28a). KG-GNN also outperformed other graph learning methods, achieving a high AUC score (Figure A.26c) and recent graph-embedding methods (Figure A.28b). These results highlight the importance of modeling the complex relationships between cell types and genes for disease phenotype prediction. We further evaluated the key components of the KG-GNN model through ablation study (Supplementary Note 6). Varying the number of genes for GRN construction showed that larger gene sets improved held-out performance but reduced generalization, while the gene set used in the original model achieved the best overall performance (Figure A.29a). Incorporating knowledge-guided graph diffusion also improved accuracy compared to no diffusion and random diffusion models

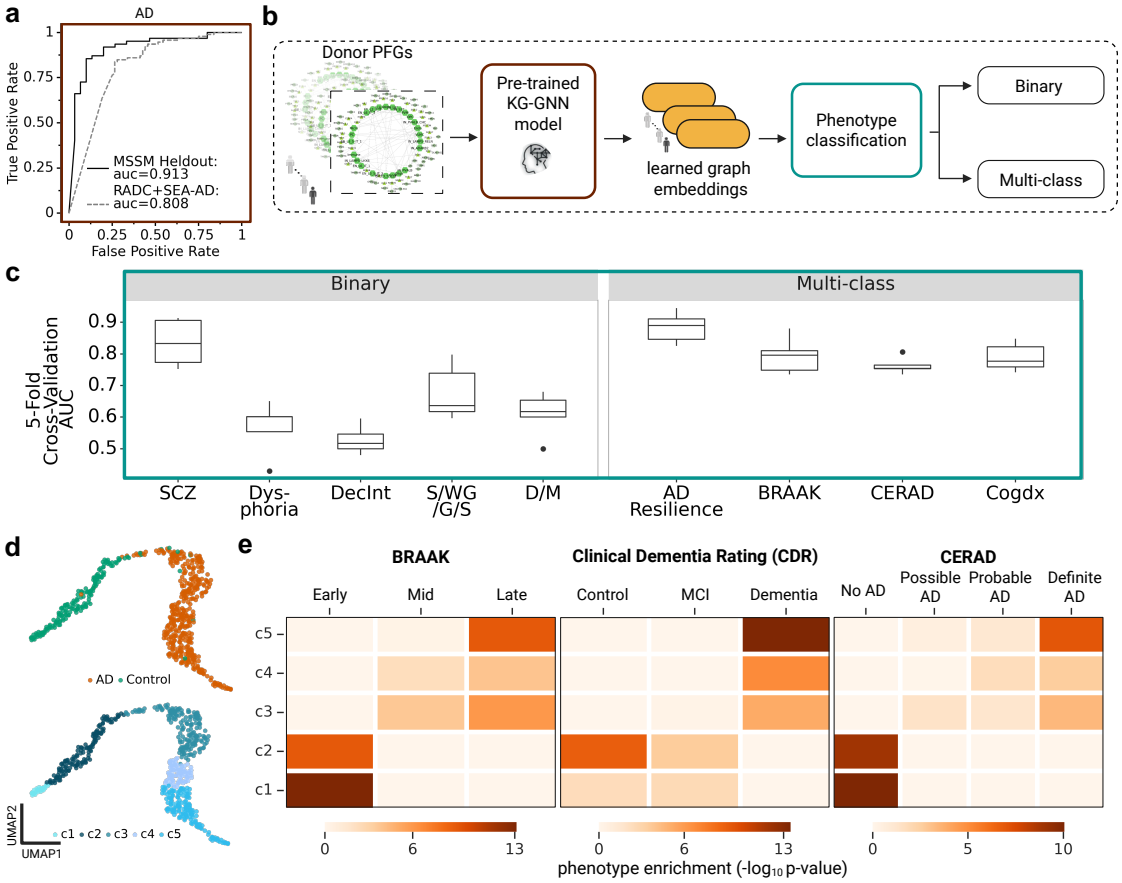

Figure 4.3: **Graph embeddings for personalized functional genomics enable phenotype classification and subtyping.** a, ROC curves for classifying AD vs. control using KG-GNN graph embeddings from the MSSM held-out: AD (n=62) vs. control (n=30) and the independent validation data RADC+SEA-AD: AD (n=93) vs. control (n=68). b, Donor PFGs from all cohorts are input into the pre-trained KG-GNN model to extract learned graph embeddings, which are then used for classifying additional phenotypes, either as binary or multi-class classification tasks. c, Average five-fold cross-validated ROC for classification of KG-GNN graph embeddings across phenotype contrasts; DecInt: Anhedonia, S/WG/G/S: Sleep/Weight Gain/Guilt/Suicide, D/M: Depression/Mood. d, Graph embeddings based UMAP (Left: colored by AD vs. controls; Right: colored by clusters from unsupervised clustering/subtyping of graph embeddings). These embeddings reflect personalized patterns derived from functional genomic graphs, capturing latent biological variation not evident in bulk or pooled analyses. e, Heatmaps depict phenotype enrichments for clusters from (c) across AD progressions: AD pathology (BRAAK and CERAD) and cognitive status (clinical dementia rating), showing an increasing trend from c1-c5, using hypergeometric test. All dataset details are described in Figure 4.2c and Supplementary Note 3.

(Figure A.29b). Finally, compared to gene co-expression networks constructed using Dozer[104], our PFGs consistently showed higher performance on both held-out and independent datasets (Figure A.29c).

Next, we applied the pre-trained KG-GNN model to PFGs from donors spanning different phenotypes beyond AD vs. control to extract graph embeddings (Figure 4.3b, Figure 4.2c). We found that these embeddings achieved good classification

performance on additional phenotypes and NPSs, using standard machine learning models for both binary (disease vs. controls and NPSs) and multi-class tasks (disease progression and AD-resilience) (Figure 4.3c, Supplementary Note 7). For binary classification, SCZ showed high performance (AUC = 0.83), while NPS contrasts such as S/WG/G/S (AUC = 0.67) and D/M (AUC = 0.61) achieved moderate accuracy. Other NPS contrasts yielded relatively lower yet comparable results. For multi-class classification, performance was above baseline across all four contrasts, including AD-resilience (AUC = 0.87), BRAAK (AUC = 0.79), CERAD (AUC = 0.75), and Cogdx (AUC = 0.79), though not as high as AD classification (Figure 4.3c, Figure A.30a). We also evaluated SCZ classification on an independent cohort from HBCC (SCZ: n=50 vs. Control: n=250), where performance was lower than the earlier AD and SCZ classifiers, suggesting the need for training disease-specific models (Figure A.31, Discussion).

We further explored potential population subtypes using our graph embeddings through unsupervised clustering and identified five distinct clusters (c1-c5, Figure 4.3d). These clusters showed clear trends across AD-related phenotypes, including BRAAK stage, cognition (Clinical Dementia Rating (CDRscore), and CERAD (Figure 4.3e). For example, c1-c2 were enriched for early BRAAK stages (0-2), controls in CDRscore, and 'No AD' based on CERAD; while c4-c5 were enriched for late BRAAK stages (5-6), Dementia (CDRscore), and presence of AD (CERAD). This suggests that our pre-trained model can identify meaningful donor subpopulations beyond AD vs. control. To assess the robustness of these trends, we varied the number of clusters and observed consistent phenotype enrichment patterns across different settings, supported by clustering quality metrics such as Davies-Bouldin and Silhouette scores (Figure A.32).

Phenotypic population trajectories for AD progression, cognition, and neuropsychiatric symptoms

AD progression exhibits heterogeneity across donors, particularly in age of onset, rate of cognitive decline, and worsening of disease pathology[15, 50, 106]. Capturing personalized disease mechanisms can therefore provide insights into disease trajectories at a population level. In this study, we used donor graph embeddings, extracted from our pre-trained KG-GNN model, to infer population-level trajectories for AD phenotypes by computing phenotypic pseudotimes across donors (Figure 4.4a, Methods). Specifically, these graph embeddings allowed us to temporally order donors by assigning pseudotimes based on their phenotypes (e.g., AD, BRAAK, CERAD). We first inferred a trajectory for all donors from the MSSM AD vs. control contrast, assigning an AD pseudotime to each donor (Figure 4.4b). AD donors were assigned later pseudotimes compared to the controls ($p < 3.17e^{-3}$, two-sided Mann-Whitney, Figure 4.4c). Furthermore, when comparing this AD pseudotime against Plaque and CDR using ordinary least squares (OLS), we observed a clear progression trend, with increasing plaque levels (r2 = 0.684, $p < 7.01e^{-106}$) and higher CDRscores (r2 = 0.856, $p < 1.81e^{-181}$) associated with later pseudotimes (Figure 4.4d). To further

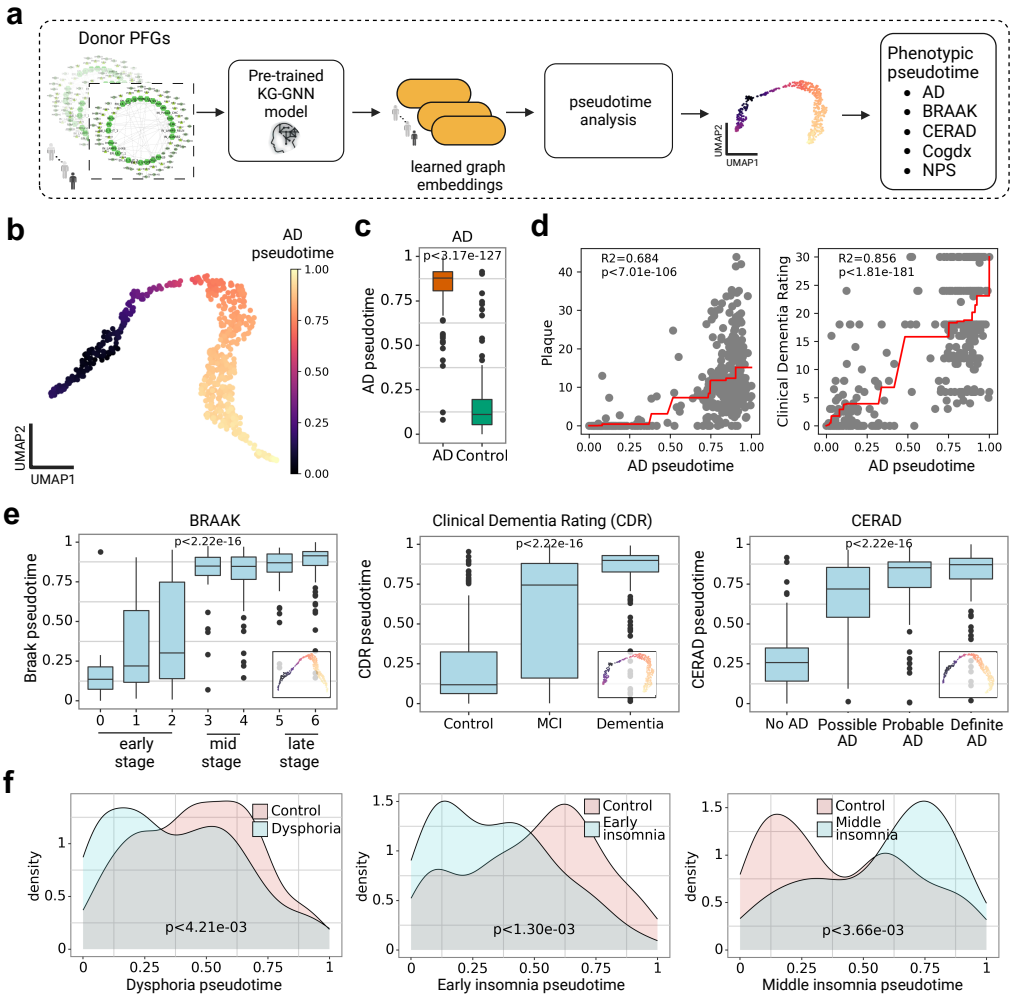

Figure 4.4: **Population-level pseudotime analysis uncovers phenotypic pseudotimes for AD progression, cognition, and NPS using pre-trained KG-GNN model.** a, Graph embeddings were extracted for donors with different phenotypic information using the pre-trained KG-GNN model. Then, pseudotime analysis of graph embeddings is performed to compute phenotypic pseudotimes, b, AD phenotypic trajectory captured for graph embeddings of donors with AD vs. controls, c, Boxplot comparing ring pseudotimes from b. shows controls appearing earlier compared to donors with AD having a later occurrence ($p < 3.17e^{-127}$, two-sided Mann-Whitney), d, Regression plots comparing AD trajectory pseudotimes (from a.) with AD pathology (Plaque) (r2 = 0.684, $p < 7.01e^{-106}$) and cognition (Clinical Dementia Rating score) (r2 = 0.856, $p < 1.81e^{-181}$), computed using ordinary least squares (OLS), e, Boxplots showing pseudotimes computed from graph embeddings for donors with AD pathology and cognition reveal an increasing trend with stage progression; left: BRAAK stages = early vs. mid vs. late ($p < 2.22e^{-16}$, two-sided Mann-Kendall), middle: CDRScore = Control vs. Mild Cognitive Impairment (MCI) vs. Dementia ($p < 2.22e^{-16}$, two-sided Mann-Kendall), right: CERAD=No AD vs. Possible vs. Probable vs. Definite AD ($p < 2.22e^{-16}$, two-sided Mann-Kendall); inset plots: graph embeddings based UMAPs colored by pseudotime. f, Density plots showing the distribution of neuropsychiatric symptoms (NPS) across donors diagnosed with NPS, including Dysphoria ($p < 4.21e^{-3}$), Early Insomnia ($p < 1.30e^{-3}$), and Middle Insomnia ($p < 4.21e^{-3}$) across NPS phenotypic pseudotimes.

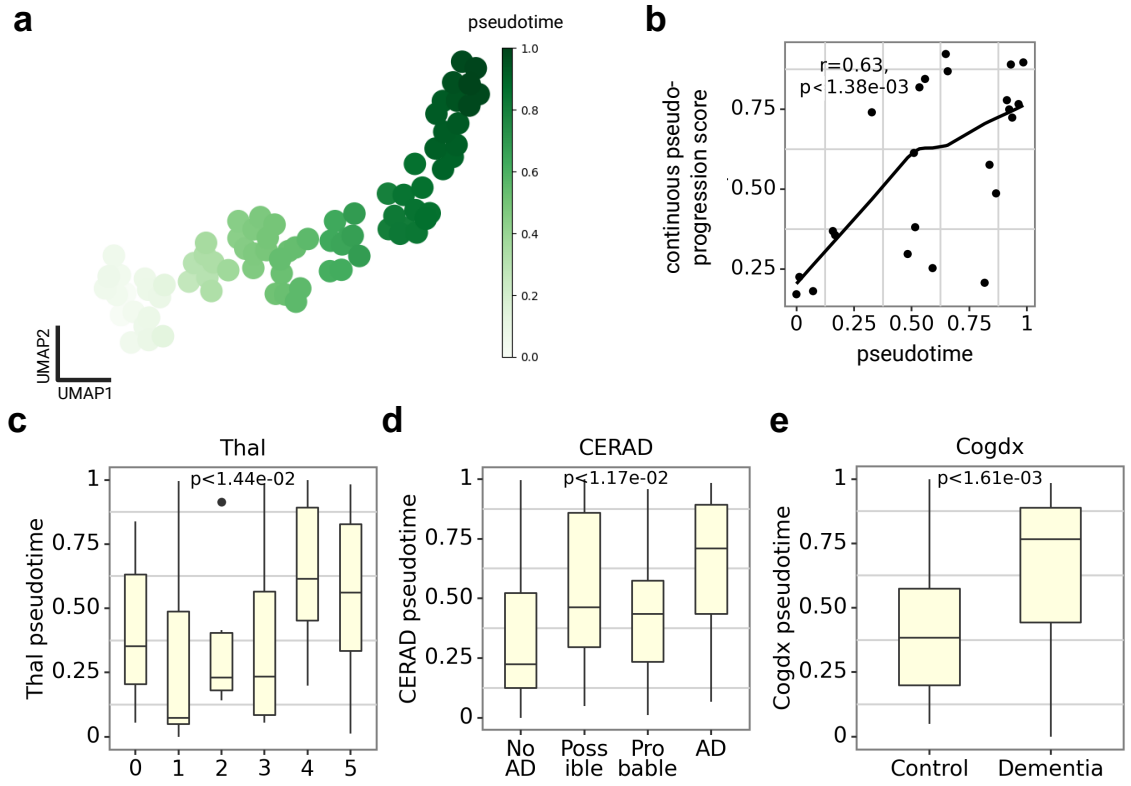

Figure 4.5: **Independent validation of phenotypic trajectories by SEA-AD cohort.** a, Graph embeddings were extracted for donors with diverse phenotypes using a pre-trained KG-GNN model to compute phenotypic pseudotimes. b, Kernel density estimate (KDE) plot comparing graph embedding based pseudotime with Continuous Pseudoprogression Score (Jensen-Shannon similarity = 0.77). c-e, Boxplots showing increasing pseudotime from (a) along AD pathology for Thal (p < 1.44e-2, two-sided Mann-Kendall) and CERAD ($p < 1.17e^{-2}$, two-sided Mann-Kendall), and cognition (Cogdx) ($p < 1.61e^{-2}$, two-sided Mann-Kendall).

explore AD phenotypic pseudotimes, we inferred trajectories for progression and cognition contrasts like BRAAK, CERAD, and CDRScore (Figure 4.4e). BRAAK staging scores range from 0-6 and were grouped into three stages: early (0-2), mid (3-4), and late (5-6). Further details of the grouping are available in Supplementary Note 3. Across all three phenotypes, we observed a consistent trend of increasing pseudotime associated with more advanced stages of disease severity ($p < 2.22e^{-16}$, $p < 2.22e^{-16}$, $p < 2.22e^{-16}$, two-sided Mann-Kendall, respectively). This suggests that our donor graph embeddings can effectively capture phenotypic trajectories, as shown by the alignment of increasing AD pseudotime with more advanced stages of AD.

NPSs often co-occur in AD donors, with different symptoms manifesting at distinct stages of the disease[91]. For instance, dysphoria is reported to be more prevalent during the early stages of cognitive impairment[91, 99]. Consistent with previous findings, our analysis revealed that pseudotimes inferred from graph embeddings of

AD donors diagnosed with dysphoria were closer to zero or initial stages of AD progression (Figure 4.4f: left). Similarly, pseudotime analysis of AD donors diagnosed with early- and mid-insomnia revealed that these sleep disturbances occur along different stages as AD progresses (Figure 4.4f: mid, right, Figure A.32).

We validated our phenotypic trajectories by extending our analysis to the sufficiently large independent SEA-AD dataset. We first constructed the bio-diffused PFGs for 80 donors from SEA-AD and extracted their graph embeddings using our pre-trained KG-GNN model (Figure 4.5a). We then compared the pseudotimes inferred from these embeddings with the continuous pseudoprogression score (CPS)[52], which used a machine-learning model on quantitative neuropathology measurements and immunohistochemical stains. Interestingly, the two measures had strong concordance, with a Jensen-Shannon similarity of 0.77, defined as 1 – Jensen-Shannon Distance[110] (Figure 4.5b). Furthermore, the pseudotimes effectively captured disease severity stages in AD progression (Thal[144], CERAD) and cognition (Cogdx) (Figure 4.5c-e).

Donor-level prioritization of cell type interactions, genes, and regulatory networks for AD

Graph embeddings improved phenotype classification and provided insights into disease progression (e.g., population trajectories). To further understand how the cell types, genes, and their interactions drive phenotype classification, we calculated the importance scores of nodes and edges for each donor (Methods). The edge importance scores are defined by the attention weights learned by the model for each cell type interaction and gene regulatory link (TF to TG). The node importance scores are derived using the edge importance scores for each gene and cell type (Supplementary Note 8). These importance scores indicate the contribution of nodes and edges to AD vs. control classification. Compared with gene expression, we observed a significantly higher correlation between our importance score and various clinical phenotypes (Table A.36). For example, gene expression of the GAD2 gene in IN_VIP cell type was not correlated with PRS while importance scores identified a positive correlation among AD donors and a negative correlation among control donors (Figure A.38). Both GAD2 and IN_VIP have been reported to be downregulated in AD in previous studies[154].

We compared edge importance scores (based on AD prior, SCZ prior, and data-driven) of cell type gene regulatory links between AD and control donors. While some significantly differential edges were consistently prioritized across all three priors, several edges were uniquely prioritized by each prior (Figure 4.6a). For example, TF IRF8 regulatory interactions in microglia had significantly higher data-driven importance scores in controls (IRF8 ß RASGEF1C: $p < 7e^{-3}$; IRF8 ß LILRB1: $p < 9e^{-4}$) but not for AD-prior and SCZ-prior. Changes in IRF8, a crucial microglial TF for AD, can cause abnormal expression of many AD-related genes[165]. Similarly, some of the interactions for ZEB1 TF (ZEB1 ß SOX5; OPC: ZEB1 ß NTNG1) were only significantly differential for SCZ-prior. Several other interac-

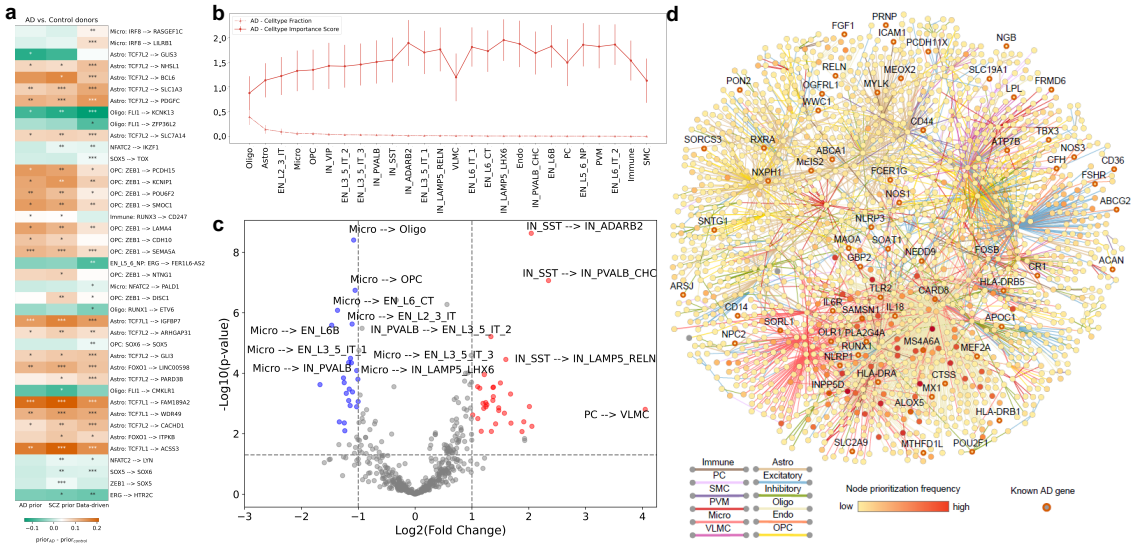

Figure 4.6: **Donor-level prioritization of cell type interactions, genes, and regulatory networks for AD.** a, Heatmap showing differences in edge importance scores for different gene regulatory links (TF-TG) across AD and control donors (annotated for significance; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$) from different biological priors: AD, SCZ, and data-driven. b, Plot showing cell type importance for AD phenotype. Dotted lines represent cell type fraction, and solid lines indicate cell type importance score. The plot is sorted by cell type fraction. The vertical lines represent the error bars. c, Volcano plot of log-fold change against significance for cell type interactions. Blue: control; red: AD. d, Prioritized subnetwork (top 1% edges across cell types within each donor) showcasing connectivity among cell types, TFs, and target genes for AD phenotype. Each circle represents a gene, shaded along a yellow to red gradient corresponding to the frequency of occurrence in prioritized nodes across donors. Edges represent regulatory links between TFs and target genes and are colored uniquely for each cell type as shown in the key below. Borders of genes linked to AD in the literature are shaded orange and labeled. Names of all other genes are hidden for ease of visualization.

tions, ZEB1 (OPC: ZEB1 ß SEMA5A; OPC: ZEB1 ß LAMA4) and TCF7L2 (Astro: TCF7L2 ß BCL6; Astro: TCF7L2 ß PDGFC), were consistently differential across all three importance scores. ZEB1, which had higher importance scores in AD donors, plays a major role in epigenetic regulation in AD[7]. TCF7L2 is associated with the Wnt/β-catenin signaling pathway which plays a crucial role in AD[141].

We examined which cell types were prioritized among AD donors and found that neuronal subtypes consistently had the highest importance scores (Figure 4.6b). For example, EN_L5_6_NP, EN_L6_IT_2, and IN_LAMP5_LHX6 were the top three cell types with the highest importance scores among AD donors. Previous transcriptomic studies have shown that the intratelencephalic neuron (IT) cell type is associated with cognitive resilience in AD[30]. The Lamp5 LHX6 inhibitory neuron (EN_LAMP5_LHX6) cell type is affected by aging and associated with AD[49, 162]. We also observed the importance scores prioritized cell types even when they had relatively low cell fractions. Furthermore, differential analysis of the edge importance scores for cell type interactions identified several AD-associated cell type

interactions. Notably, interactions originating from microglia were higher in control donors, whereas those involving IN_SST neurons as the source cell type were higher in AD donors (Figure 4.6c). Microglia cell type is known to have many effects on AD pathology and therapeutic intervention, which corroborates their high importance[65]. Somatostatin inhibitory neurons (IN_SST) are also linked to memory loss and SST expression is reduced among AD patients[137].

We then combined the top-prioritized regulatory links from each donor (top 1% based on our edge importance scores) into a consensus functional genomic subnetwork (Figure 4.6d). The subnetwork segregated the edges into three main clusters consisting mainly of oligodendrocytes, astrocytes, and excitatory neuronal cell types (Figure A.33), indicating that our model consistently prioritized regulatory links in these cells across donors. Oligodendrocytes are responsible for myelination, protecting neurons from damage, and dysfunction in oligodendrocytes leads to demyelination, a major factor in AD progression[41, 132]. Astrocytes play a critical role in AD pathogenesis by contributing to neuroinflammation, synaptic dysfunction, and amyloid-β ($A\beta$) plaque clearance[127, 60, 135]. Excitatory neurons are particularly vulnerable in AD, exhibiting early synaptic dysfunction, loss of dendritic spines, and impaired glutamatergic signaling, which contribute to cognitive decline[116]. Other single-cell transcriptomic analyses revealed distinct molecular alterations in excitatory neuron subtypes in AD[108], astrocyte subclusters exhibiting varying response to neuropathology in AD progression[107, 58], and AD-associated oligodendrocyte subpopulations with altered gene expression related to myelination and cellular responses[169]. This subnetwork recapitulates several biologically relevant features of our model, with a statistically significant overrepresentation of known AD genes ($p < 9.80e^{-4}$, hypergeometric test). For instance, PCDH11X, an aging-related gene associated with late-onset AD in GWAS[21], is prioritized in excitatory neurons in our model. Cross-verifying with an external independent dataset[44] confirms PCDH11X predominant expression in excitatory neurons (Figure A.34). A previous study identified variants in the CR1 gene associated with AD[90]. Our model suggests that FOSB mediates CR1 regulation, and this regulation is relevant for AD PVM cells. Overall, our model effectively prioritized a subnetwork that recapitulates known AD biology and suggests cell-type-specific regulation of several genes with prior evidence. Thus, other genes in the prioritized subnetwork are likely candidates that have remained elusive thus far.

Imputing graph embeddings using genotype

We also explored whether genotype data can be used to impute graph embeddings and whether these imputed embeddings could classify donor phenotypes (Figure A.35a, Supplementary Note 9). We trained our imputation pipeline using the PsychAD data and validated the phenotype classification on two independent datasets, ROSMAP[38] and ADNI[119], where only genotype data was available. In the ROSMAP cohort, we classified donors into early vs. late BRAAK stages, while in ADNI, we classified donors into AD vs. control. We evaluated several cross-modality

imputation methods (CMOT[4], JAMIE[29], MOFA+[8], and autoencoders) and observed that the best performing model achieved an AUC of 0.57 for phenotype classification on both cohorts. Although modest, this performance was above random and comparable to results reported in prior PRS studies[96] (Figure A.35b, Table A.35). We next clustered the imputed donor graph embeddings and identified seven potential subtypes (Figure A.35c). We found that clusters c2-c6 were enriched with CERAD stages, which progressed in severity from 'No AD' to 'Definite' (Figure A.35d). These findings highlight how SNPs relate to cell type gene regulation changes and open up new possibilities for using genotype data to classify disease phenotypes and discover subtypes, even when transcriptomic data is not available.

4.4 Methods

Datasets and Preprocessing

Datasets

Dataset Our analysis is focused on the population-level snRNA-seq data from the PsychAD consortium covering the dorsolateral prefrontal cortex (DLPFC) brain region from 1494 donors, derived from three cohorts: NIH NeuroBioBank at the Mount Sinai Brain Bank/MSSM (M), Rush Alzheimer's Disease Center/RADC (R), and the NIMH-IRP Human Brain Collection Core/HBCC (H). The PsychAD data description is available in the "PsychAD dataset" (Supplementary Information). Our analyses utilized disease diagnosis (e.g., AD, SCZ, DLBD), quantitative assessment of the disease stages for donors with AD (e.g., BRAAK or CERAD), and diagnosis of neuropsychiatric symptoms (e.g., depression or weight loss). Accordingly, these donors were categorized into three levels: disease vs. control (AD, SCZ, DLBD, AD Resilience), AD progression (BRAAK stages, Cogdx, CERAD), and NPSs (Dysphoria, DecInt (Anhedonia), Sleep/Weight Gain/Guilt/Suicide (S/W/G/S), Depression/Mood (D/M)) (Figure 4.2, Supplementary Note 3). We mainly used the MSSM (M) cohort for all the training and used RADC (R) as well as an external dataset from The Seattle Alzheimer's Disease Brain Cell Atlas/SEA-AD (S)[68] for independent validation.

Data processing and feature selection

PsychAD snRNA-seq data was preprocessed as described in Lee et al.[92] and Fullard et al.[51]. The preprocessing involved identifying 5,000 highly variable genes (HVGs). We selected the protein-coding genes from these 5,000 HVGs and intersected them with the genes from the SEA-AD dataset. We ended up with 2,766 HVGs which were used as features in our analysis. We applied standard Scanpy (v1.9.3) functions to preprocess SEA-AD data. For SEA-AD, we selected DLPFC regions of 39 donors with dementia and 39 without dementia to retain individual balance. For independent validation, we constructed bio-diffused PFGs for the donors and included the same set of 2,766 HVG genes as features as the PsychAD dataset.

Construction of personalized functional genomics graph

For each donor, we define his/her personalized functional genomics graph (PFG) as a directed graph with two major node types: cell types and genes. Gene nodes include cell-type-specific transcription factors (TFs) and target genes (TGs). Each edge in the PFG conveys distinct information depending on the type of nodes. For example, cell type to cell type edge captures cell type interactions, TF-TG edge captures gene regulatory links, and cell type-TF edge captures cell-type-specific relationship of TFs. As some of the donors had fewer cells for the cell type interactions and gene regulatory network algorithms to capture (Figure A.37, we were able to construct PFGs for 1478 out of 1494 donors. Below, we provide details on constructing each component of PFG.

Inferring cell type interactions

For each donor, the cell type to cell type links were constructed using CellChat[82] to quantify the cell-cell communications. CellChat provides a consolidated interaction probability for each cell type to cell type interaction by adding all the probabilities of ligand-receptor interactions between two cell types. We used snRNA-seq data from each donor as input to CellChat with 27 subclass information as the label. We used the default settings in CellChat to extract the probabilities using the function computeCommunProb. As the consolidated interaction probability can exceed 1, we normalized the probability between 0 and 1 and retained all links with the normalized score above a 0.5 threshold to construct cell type to cell type links for each donor.

Extracting cell type gene regulatory links

We first used GRNBoost2 on the snRNA-seq data for each donor to deduce gene co-expression networks across all cell types. SCENIC[3] was then used to infer cell type gene regulatory links. Specifically, we applied runscENIC_1_coexNetwork2modules and runscENIC_2_createRegulons functions in Python with default settings (constraining the TF search to 10 kbps radius around the TSS or 500 bp upstream) to identify regulons. Regulons with at least 10 genes were scored in each cell using runscENIC_3_scoreCells. We then computed AUCell enrichment based on the top 1% of genes detected per cell. We incorporated the regulon specificity score (RSS)[140] to evaluate the cell-type-specificity of the regulon. For each cell type, we select the top 10% of the regulons based on the RSS score as the cell type regulons. We then keep the top 10% of the TGs for each cell type regulon. Together, these comprise the cell type gene regulatory links for each donor.

Definition of PFGs

Let $G_i = (V_i, E_i)$ (Figure 4.2a) be a PFG for donor i, with nodes and representing the edges. Then, an edge $e_{i,st} \in E_i$ connecting the source node s to a target node t represents one of the following relationships:

$$e_{i,st} = \begin{cases} 1, & \text{if s interacts with t, where s \& t are cell types,} \\ 1, & \text{if s regulates with t, where s is a TF, t is a TG,} \\ 1, & \text{if t belongs to the cell type regulon of s, where s is a cell type, t is TF,} \\ 1, & \text{if t belongs to the cell type regulon of s, where s is a cell type, t is TG,} \\ 0, & \text{otherwise,} \end{cases} \tag{4.1}$$

The node features used for each PFG depend on the type of node. We used the average gene expression of 2,766 highly variable genes (HVGs) from all cells as the features for cell type (CT) nodes and the co-expression of these HVGs with the gene expression for the TF/TG nodes.

Integrating prior biological knowledge into PFGs via network diffusion

We then applied network diffusion to propagate the influence of known disease genes on the PFGs (Supplementary Note 1). This diffusion process allows us to incorporate known disease-related information into our graph structure, potentially enhancing the model's ability to capture disease-relevant patterns. We currently use well-known disease genes for AD and SCZ extracted from the DisGeNet database[121] and filtered based on manually curated sources with "CTD_human" or "GWAS-CAT" identifiers. Additionally, we downloaded high-confidence SCZ genes from the PsychENCODE project[44]. In total, we identified 361 AD genes and 945 SCZ genes. The resulting matrices are called "bio-diffused PFGs" and are used to train our graph neural network model. While we focus on AD and SCZ in this paper, this method can be applied to any gene-of-interest (GOI) list.

Graph classification using Knowledge-guided Graph Neural Network (KG-GNN)

The KG-GNN model is based on the multi-head Graph Attention Networks (GAT)[151] that learns from arbitrary bio-diffused PFGs as inputs for classifying AD vs. control. It is an interpretable machine learning model that integrates prior biological knowledge, personalized functional genomics, and biologically driven multi-head graph attention networks to derive insights about disease and underlying disease mechanisms at a donor level. The details of training procedure, validation, and benchmarking are provided in Supplementary Note 4-5. The input to the KG-GNN mode is the PFGs. The trained model outputs graph embeddings and personalized AD importance scores for nodes and edges.

Population subtyping and enrichment based on graph embeddings

We clustered graph embeddings extracted for donors with AD vs. controls using the function sc.tl.louvain() from the Python package scanpy[160]. Before clustering, we identified the nearest neighbors using the sc.pp.neighbors() function and computed

the UMAP embeddings using the sc.tl.umap() function. Then we performed Louvain clustering using the function sc.tl.louvain() and set the resolution = 0.35 to get 5 clusters. We clustered the graph embeddings to identify potential novel subtypes. To identify novel subtypes, we performed enrichment analysis using a hypergeometric test to see if any clusters were enriched with AD progression stages. We used the stats.hypergeom.cdf() from the Python package Scipy[153] for this analysis and reported the -log10(p-value) of phenotype enrichment in each cluster.

Phenotypic population trajectory inference

We inferred phenotypic trajectories across AD phenotypes based on donor graph embeddings extracted from our pre-trained KG-GNN model. Here we used a diffusion-based algorithm[63] within the Scanpy package in Python to infer trajectories. Pseudotime was computed using sc.tl.diffmap() and sc.tl.dpt() functions from the Scanpy package. To determine phenotypic pseudotime for donors with a given phenotype, we selected a starting point or root by randomly choosing a donor from those assigned to the earliest phenotype stages. For AD & BRAAK, this was a donor from stage 0; for CERAD, a donor with a CERAD score of 0 (No AD); and for Cogdx, a donor with a cognitive diagnosis of 1 (control). To infer the pseudotimes for different NPS, we again set a random donor with a CERAD score of 2 as the root. This is because all donors diagnosed with NPS have a CERAD score equal to at least 2. To infer the trajectory for SEA-AD donors, we first extracted their graph embeddings from the pre-trained KG-GNN model. Then using similar functions as above (sc.tl.diffmap(), sc.tl.dpt()), we computed a trajectory by randomly selecting a donor with CERAD score = "Absent" as the root.

Calculation of importance scores

The trained KG-GNN model outputs edge attentions which are used as edge importance scores. We get three different edge importance scores: AD-prior, SCZ-prior, and data-driven based on prior biological knowledge from known AD-genes and SCZ-genes. We derive the node importance scores using the edge importance scores. We use a combination of both incoming and outgoing edge importance as the basis to calculate these scores. Further details about node importance calculation are available in Supplementary Note 8. We averaged the importance scores across all three priors for each node and edge per donor.

Chapter 5

Discussion

CMOT is a computational approach that integrates manifold alignment, regularized optimal transport, and k nearest neighbors for cross modality inference. By applying emerging single cell multimodal data, we demonstrated that CMOT can predict multimodal features of single cells, including gene expression, chromatin accessibility, and protein abundance. Notably, CMOT does not require paired samples for aligning modalities, and outperformed existing methods in several scenarios. This is especially useful since joint multimodal profiling is often difficult and expensive, and single modality data remains more common. For instance, when evaluated on separate scRNAseq and scATACseq profiles from human brain data [149], CMOT surpassed all other methods in performance (Figure A.14) and was computationally more efficient (Table A.29). While the method primarily uses nonlinear manifold alignment, CMOT is modular and allows for alternative alignment strategies such as SCOT [39], MMD MA [100], or WNN [66].

The optimal transport step in CMOT quantifies the difference between source and target distributions using Wasserstein distances and computes a mapping that minimizes this discrepancy, leading to better cross modality inference. However, CMOT has some limitations. Nonlinear manifold alignment becomes computationally expensive as data size grows, due to the need to compute similarity matrices for each modality. This cost could be reduced using faster alignment methods such as SCOT [39], [40] or Unioncom [18]. In addition, the mass balanced nature of traditional optimal transport assumes that every source point is mapped to some target point. This assumption may not hold in the presence of imbalance between source and target distributions. Unbalanced optimal transport approaches such as SCOTv2 [40] and related work [28], [142] can help address this issue.

CMOT can also be extended with other optimal transport variants. For example, Gromov Wasserstein distances enable mapping across unmatched modalities [6], [112]. Additionally, CMOT could be applied to infer cell labels across modalities, predicting cell types or disease states in data lacking such annotations [66]. Finally, it has the potential to incorporate newer single cell modalities such as cell morphology from Patch seq data.

Building on the goal of modeling biological processes from unpaired or sparse

observations, we next developed ARTEMIS, a generative model for reconstructing cellular trajectories and estimating the regulatory drift that drives them. ARTEMIS learns interpretable, low dimensional representations of single cell gene expression data using a variational autoencoder and combines this with unbalanced Schrödinger bridges to capture population level changes over time [86].

ARTEMIS performs well on sparse time series data (Appendix A.2 (Supplementary Note S7), Table A.33). However, the lack of intermediate timepoints may reduce resolution in trajectory inference. Technical variation across experiments may also affect results, although the VAE helps mitigate this by introducing latent randomness, reducing overfitting [147]. In our case, batch effects were either corrected prior to modeling or could be addressed using tools such as Seurat [67] or Scanpy [160].

Joint training of the VAE and Schrödinger bridge introduces additional complexity. We evaluated model performance across different hyperparameter settings (Appendix A.2 (Supplementary Note S4), Figure A.18) to offer heuristic guidance. While ARTEMIS captures trends in population dynamics, predicting exact cell states such as birth or death at missing timepoints remains difficult. Future extensions of the model could incorporate prior knowledge, such as gene sets related to cell cycle or apoptosis, to enhance biological relevance. We also plan to incorporate other modalities such as chromatin accessibility to better understand the regulatory mechanisms underlying dynamic processes.

Finally, we expanded our analysis to population scale variation by introducing iBrainMap, a framework for personalized functional genomics in Alzheimer's disease using single nucleus RNA sequencing. iBrainMap constructs individual specific functional graphs that capture transcriptional regulation at the donor level. Using these graphs, we trained a knowledge guided graph neural network model to classify phenotypes, identify novel disease subtypes, and stratify donors. Graph embeddings extracted from the trained model were used to classify several AD related traits. We validated these results on external datasets, including ROSMAP and SEA AD, demonstrating generalizability. The framework is available through an interactive web interface and can be applied to other complex diseases such as schizophrenia, bipolar disorder, or cancer when similar data are available.

By modeling inter individual differences in cell type specific regulatory activity, iBrainMap can potentially support the identification of personalized therapeutic targets. Prior work has shown that cell type specific networks can improve prediction of drug targets and clinical phenotypes in AD [59]. For instance, donepezil was found to have beneficial effects in female patients with mild cognitive impairment carrying the BCHE K variant [36], and bumetanide has been identified as a potential treatment for APOE $\varepsilon 4$ carriers through computational drug screening [143]. Our approach can be extended to other genetically diverse diseases including schizophrenia, autism spectrum disorder [44], and cancer [71].

Several limitations remain. The regulatory graphs were constructed using a fixed threshold on RSS scores, which may miss subtle interactions. Future efforts could expand the graph structure by incorporating other biological networks such as pro-

tein interaction maps. Our graph neural network model currently treats all nodes and edges uniformly, which may overlook distinctions between transcription factors, target genes, and cell types. More expressive models such as heterogeneous graph neural networks [105] may offer improvements. Classification performance was lower for comorbid traits such as neuropsychiatric symptoms (Figure A.27b) and AD with Lewy body disease (Figure A.30b), likely because the model was primarily trained on AD versus control. Similarly, importance scores for schizophrenia related graphs showed higher variance (Figure A.36). This suggests that training disease specific models when sufficient data are available may improve results. We also evaluated a genotype based model for imputing donor embeddings, which could be used to make phenotype predictions in the absence of expression data. While performance was modest, it was in line with previous studies using polygenic risk scores [96]. More accurate prediction of disease states may require integration with additional data such as epigenomics, transcriptomics, or environmental exposures [80, 111].

Chapter 6

Conclusion

This thesis presents a suite of computational frameworks designed to advance our ability to analyze and interpret high-dimensional single-cell data, with a focus on cross-modality inference, dynamic trajectory modeling, and personalized functional genomics.

In the first part of this work, I introduced CMOT, a cross-modality inference framework that leverages nonlinear manifold alignment and regularized optimal transport to align cells across modalities and infer missing molecular features. CMOT addresses the pervasive challenge of incomplete multimodal data by enabling joint analysis even in the absence of paired measurements. It demonstrates improved performance across a range of biological systems, from brain development to cancer and immunology, and offers a flexible and computationally efficient foundation for cross-modal data integration.

In the second part, I developed ARTEMIS, a generative model that reconstructs cellular trajectories and captures population dynamics from time-series single-cell transcriptomic data. By combining a variational autoencoder with an unbalanced Schrödinger bridge, ARTEMIS models both stochastic and deterministic aspects of cell state transitions while estimating birth-death processes over time. It outperforms existing methods in reconstructing trajectories and identifying genes driving deterministic trends, offering insights into the temporal dynamics of development and disease.

Finally, I presented iBrainMap, a personalized functional genomics framework applied to Alzheimer's disease. By constructing donor-specific functional genomic graphs and applying knowledge-guided graph neural networks, this approach preserves inter-individual variation and enables stratification of disease phenotypes. iBrainMap captures donor-level differences in cell types and regulatory interactions, enhances phenotype classification, and reveals biologically meaningful subtypes in complex brain disorders. It serves as a publicly available resource and establishes a scalable strategy for population-scale functional genomic modeling.

Together, these contributions demonstrate how integrative machine learning methods can address key challenges in single-cell biology, including missing modalities, temporal alignment, and donor-level heterogeneity. While each framework was de-

veloped for a specific task, they collectively highlight the importance of modeling biological complexity through principled, data-driven approaches. Future directions include extending these models to incorporate multimodal and spatial data, improving computational scalability, and applying them to additional diseases and tissues. This work lays the foundation for developing more refined, individualized models of cellular function that will be critical for advancing both basic biology and precision medicine.

Appendix A

Appendix

A.1 CMOT: Supplement

Supplementary Tables

Table A.1: **Gene expression inference from chromatin accessibility in human developing brain data [149]**. Mean and median cell-wise Pearson correlation between inferred and measured gene expression comparing methods CMOT(p=25%,50%,75%,100%), Seurat, MOFA+, GLUE, bindSC, and Seurat CCA plotted in Fig. S6 for human developing brain.

Methods	Mean Pearson Correlation	Median Pearson Correlation
CMOT (p=100%)	0.66	0.67
CMOT (p=75%)	0.65	0.68
CMOT (p=50%)	0.63	0.65
CMOT (p=25%)	0.58	0.61
Seurat	0.62	0.64
MOFA+	0.43	0.41
GLUE	0.46	0.47
bindSC	0.63	0.68
Seurat CCA	0.66	0.68

Table A.2: **Gene expression inference from chromatin accessibility in human developing brain data [149]**. Mean and median cell-wise Spearman correlation between inferred and measured gene expression comparing methods CMOT (p=25%,50%,75%,100%), Seurat, MOFA+, GLUE, bindSC, and Seurat CCA plotted in Fig. S6 for human developing brain. All values calculated with alternative=`greater`.

Methods	Mean Spearman Correlation	Median Spearman Correlation
CMOT (p=100%)	0.44	0.44
CMOT (p=75%)	0.44	0.44
CMOT (p=50%)	0.43	0.44
CMOT (p=25%)	0.43	0.43
Seurat	0.45	0.45
MOFA+	0.17	0.16
GLUE	0.42	0.42
bindSC	0.44	0.45
Seurat CCA	0.44	0.44

Table A.3: **Gene expression inference from chromatin accessibility in human developing brain data [149]**. Wilcox–rank-sum test p-values of cell-wise Pearson correlation between inferred and measured gene expression comparing methods CMOT (p=25%,50%,75%,100%), Seurat, MOFA+, GLUE, bindSC, and Seurat CCA plotted in Fig. S6 for human developing brain. All values calculated with alternative=`greater`.

Methods	Seurat	MOFA+	GLUE	bindSC	Seurat CCA
CMOT (p=100%)	1.23×10^{-14}	0	1.4×10^{-236}	0.97	0.98
CMOT (p=75%)	3.4×10^{-10}	2.8×10^{-294}	1.3×10^{-216}	0.99	0.99
CMOT (p=50%)	0.3	3.3×10^{-240}	1.5×10^{-162}	1	1
CMOT (p=25%)	1	1.65×10^{-157}	2.4×10^{-86}	1	1

Table A.4: **Gene expression inference from chromatin accessibility in human developing brain data [149]**. Wilcox–rank-sum test p-values of cell-wise Spearman correlation between inferred and measured gene expression comparing methods CMOT(p=25%,50%,75%,100%), Seurat, MOFA+, GLUE, bindSC, and Seurat CCA plotted in Table S2 for human developing brain. All values calculated with alternative=`greater`.

Methods	Seurat	MOFA+	GLUE	bindSC	Seurat CCA
CMOT (p=100%)	0.93	0	5.51×10^{-30}	0.99	0.27
CMOT (p=75%)	0.99	0	6.36×10^{-21}	1	0.97
CMOT (p=50%)	1	0	4.49×10^{-12}	1	0.99
CMOT (p=25%)	1	0	2.87×10^{-5}	1	1

Table A.5: **Gene expression inference from chromatin accessibility in mouse brain data [26]**. Mean and median cell-wise Pearson correlation between inferred and measured gene expression comparing methods CMOT(p=25%,50%,75%,100%), MOFA+, Seurat, GLUE, bindSC, and Seurat CCA plotted in Fig. S1 for mouse brain.

Methods	Mean Pearson Correlation	Median Pearson Correlation
CMOT (p=100%)	0.71	0.76
CMOT (p=75%)	0.71	0.76
CMOT (p=50%)	0.71	0.76
CMOT (p=25%)	0.67	0.71
Seurat	0.66	0.70
MOFA+	0.69	0.74
GLUE	0.67	0.72
bindSC	0.68	0.72
Seurat CCA	0.70	0.75

Table A.6: **Wilcox–rank-sum test p-values of cell-wise Spearman correlation in mouse brain data [26]**. Comparing methods CMOT(p=25%,50%,75%,100%) and competing methods under alternative=`greater`.

Methods	Mean Spearman Correlation	Median Spearman Correlation
CMOT (p=100%)	0.34	0.34
CMOT (p=75%)	0.34	0.34
CMOT (p=50%)	0.35	0.35
CMOT (p=25%)	0.35	0.35
Seurat	0.40	0.40
MOFA+	0.31	0.32
GLUE	0.32	0.33
bindSC	0.44	0.46
Seurat CCA	0.36	0.36

Table A.7: **Wilcox–rank-sum test p-values of cell-wise Pearson correlation in mouse brain data [26]**. Comparing methods CMOT(p=25%,50%,75%,100%) and MOFA+.

Methods	Seurat	MOFA+	GLUE	bindSC	Seurat CCA
CMOT (p=100%)	3.4×10^{-40}	1.15×10^{-5}	5.6×10^{-25}	8×10^{-19}	1.9×10^{-4}
CMOT (p=75%)	1.25×10^{-39}	1.67×10^{-5}	1.63×10^{-24}	1.8×10^{-18}	2.8×10^{-4}
CMOT (p=50%)	1.34×10^{-35}	8×10^{-4}	5.04×10^{-21}	2.06×10^{-15}	8×10^{-3}
CMOT (p=25%)	0.07	1	0.9	0.99	1

Table A.8: **Wilcox–rank-sum test p-values of cell-wise Spearman correlation in mouse brain data [26]**. Comparing methods CMOT(p=25%,50%,75%,100%) and MOFA+.

Methods	Seurat	MOFA+	GLUE	bindSC	Seurat CCA
CMOT (p=100%)	1	5×10^{-28}	6.1×10^{-13}	1	1
CMOT (p=75%)	1	4.72×10^{-30}	2.04×10^{-14}	1	1
CMOT (p=50%)	1	5×10^{-37}	1.64×10^{-19}	1	1
CMOT (p=25%)	1	1.47×10^{-55}	3×10^{-34}	1	1

Table A.9: **Gene expression inference from chromatin accessibility in mouse brain data [26]**. Mean and median cell-wise Pearson correlation between inferred and measured gene expression comparing methods CMOT(p=25%,50%,75%,100%), MOFA+, Seurat, Polarbear, and Polarbear-coassay plotted in Fig. S12 for mouse brain.

Methods	Mean Pearson Correlation	Median Pearson Correlation
CMOT (p=100%)	0.67	0.71
CMOT (p=75%)	0.67	0.71
CMOT (p=50%)	0.66	0.70
CMOT (p=25%)	0.64	0.68
Seurat	0.60	0.63
MOFA+	0.68	0.71
Polarbear	0.68	0.71
Polarbear-coassay	0.67	0.71

Table A.10: **Wilcox–rank-sum test p-values of cell-wise Spearman correlation in mouse brain data [26]**. Comparing methods CMOT(p=25%,50%,75%,100%), Polarbear, and MOFA+ plotted in Fig. S12 for mouse brain.

Methods	Seurat	MOFA+	Polarbear	Polarbear-coassay
CMOT (p=100%)	2.97×10^{-69}	0.7	0.7	0.29
CMOT (p=75%)	2.97×10^{-69}	0.7	0.7	0.29
CMOT (p=50%)	1.51×10^{-51}	0.9	0.9	0.9
CMOT (p=25%)	9.05×10^{-26}	1	1	1

Table A.11: **Protein expression inference from gene expression in PBMC [54]**. Mean and median cell-wise Pearson correlation between inferred and measured protein expression comparing methods CMOT(p=25%,50%,75%,100%), Seurat, MOFA+, TotalVI plotted in Fig. 3A.

Methods	Mean Pearson Correlation	Median Pearson Correlation	
CMOT (p=100%)	0.78	0.86	
CMOT (p=75%)	0.78	0.86	
CMOT (p=50%)	0.77	0.85	
CMOT (p=25%)	0.74	0.83	
Seurat	0.84	0.85	
MOFA+	0.75	0.79	
TotalVI	0.61	0.61	

Table A.12: **Wilcox–rank-sum test p-values of cell-wise Pearson correlation in PBMC [54]**. Comparing methods CMOT(p=25%,50%,75%,100%), Seurat, and MOFA+.

Methods	Seurat	MOFA+	TotalVI
CMOT (p=100%)	1	6.9×10^{-57}	0
CMOT (p=75%)	1	0	8.36×10^{-58}
CMOT (p=50%)	1	0	1.73×10^{-45}
CMOT (p=25%)	1	0	5.25×10^{-12}

Table A.13: **Pearson and Spearman correlation between inferred and measured protein expression in PBMC [54]**. Independent evaluation for CMOT.

Protein	Pearson Corr.	P-value	Spearman Corr.	P-value
CD3	0.92	0	0.79	0
CD4	0.83	0	0.76	0
CD8a	0.54	3.1×10^{-312}	0.018	0.25
CD14	0.95	0	0.67	0
CD15	0.50	1.58×10^{-252}	0.46	1.18×10^{-218}
CD16	0.75	0	0.40	3.98×10^{-153}
CD56	0.69	0	0.34	2.16×10^{-114}
CD19	0.86	0	0.23	1.92×10^{-50}
CD25	0.40	3.06×10^{-144}	0.39	2.07×10^{-145}
CD45RA	0.72	0	0.72	0
CD45RO	0.73	0	0.70	0
PD-1	0.31	2.36×10^{-89}	0.26	8.95×10^{-65}
TIGIT	0.45	6.17×10^{-197}	0.34	9.29×10^{-112}
CD127	0.85	0	0.71	0

Table A.14: **Protein expression inference from gene expression in PBMC [54]**. Mean and median cell-wise Pearson correlation comparing CMOT(p=25%,50%,75%,100%), Seurat, MOFA+, TotalVI plotted in Fig. S8.

Methods	Mean Pearson Corr.	Median Pearson Corr.
CMOT (p=100%)	0.83	0.91
CMOT (p=75%)	0.83	0.90
CMOT (p=50%)	0.82	0.90
CMOT (p=25%)	0.80	0.88
Seurat	0.87	0.92
MOFA+	0.83	0.89
TotalVI	0.01	-0.08

Table A.15: **Wilcox–rank-sum test p-values of cell-wise Pearson correlation in PBMC [54]**. Comparing CMOT(p=25%,50%,75%,100%), Seurat, and MOFA+ plotted in Fig. S8.

Methods	Seurat	MOFA+	TotalVI
CMOT (p=100%)	0.99	1.2×10^{-5}	0
CMOT (p=75%)	1	1.4×10^{-2}	0
CMOT (p=50%)	1	0.13	0
CMOT (p=25%)	1	0.99	0

Table A.16: **Pearson and Spearman correlation between inferred and measured protein expression in PBMC 10K [54]**. Independent evaluation for CMOT.

Protein	Pearson Corr.	P-value	Spearman Corr.	P-value
CD3	0.95	0	0.80	5.5×10^{-317}
CD4	0.90	0	0.86	0
CD8a	0.82	0	0.40	4.1×10^{-52}
CD14	0.96	0	0.55	1.6×10^{-111}
CD15	0.56	1×10^{-118}	0.65	8.8×10^{-169}
CD16	0.79	2×10^{-302}	0.54	7.5×10^{-107}
CD56	0.84	0	0.57	1.2×10^{-122}
CD19	0.95	0	0.18	1.34×10^{-11}
CD25	0.50	7.7×10^{-90}	0.51	2.25×10^{-95}
CD45RA	0.83	0	0.83	0
CD45RO	0.80	5.9×10^{-306}	0.76	3.6×10^{-262}
PD-1	0.43	9.6×10^{-66}	0.32	7.28×10^{-35}
TIGIT	0.72	5.7×10^{-222}	0.47	9.8×10^{-78}
CD127	0.85	2.2×10^{-16}	0.80	8.6×10^{-316}

Table A.17: **Gene expression inference from chromatin accessibility in DEX-treated A549 lung cancer data [17]**. Mean and median cell-wise Pearson correlation comparing CMOT(p=25%,50%,75%,100%), Seurat, MOFA+, GLUE, bindSC, Seurat CCA plotted in Fig. S7.

Methods	Mean Pearson Corr.	Median Pearson Corr.
CMOT (p=100%)	0.52	0.52
CMOT (p=75%)	0.51	0.52
CMOT (p=50%)	0.50	0.51
CMOT (p=25%)	0.49	0.50
Seurat	0.49	0.50
MOFA+	0.52	0.52
GLUE	0.49	0.50
bindSC	0.40	0.51
Seurat CCA	0.50	0.51

Table A.18: **Gene expression inference from chromatin accessibility in DEX-treated A549 lung cancer data [17]**. Mean and median cell-wise Spearman correlation comparing CMOT(p=25%,50%,75%,100%), Seurat, MOFA+, GLUE, bindSC, Seurat CCA plotted in Fig. S7. All values calculated with alternative=`greater`.

Methods	Mean Spearman Corr.	Median Spearman Corr.
CMOT (p=100%)	0.43	0.43
CMOT (p=75%)	0.43	0.43
CMOT (p=50%)	0.43	0.43
CMOT (p=25%)	0.42	0.42
Seurat	0.41	0.41
MOFA+	0.44	0.44
GLUE	0.43	0.44
bindSC	0.39	0.49
Seurat CCA	0.42	0.42

Table A.19: **DEX-treated A549 lung cancer (Pearson)** [17]. Wilcox–rank-sum test p-values comparing CMOT, Seurat, MOFA+, GLUE, bindSC, Seurat CCA (Fig. S7).

Methods	Seurat	MOFA+	GLUE	bindSC	Seurat CCA
CMOT (p=100%)	1.27×10^{-5}	0.64	4.7×10^{-6}	0.016	0.016
CMOT (p=75%)	3.9×10^{-5}	0.74	1.6×10^{-5}	0.025	0.029
CMOT (p=50%)	0.012	0.99	0.007	0.22	0.45
CMOT (p=25%)	0.70	0.99	0.66	0.82	0.99

Table A.20: **DEX-treated A549 lung cancer (Spearman) [17]**. Wilcox–rank-sum p-values comparing CMOT, Seurat, MOFA+, GLUE, bindSC, Seurat CCA (Fig. S7).

Methods	Seurat	MOFA+	GLUE	bindSC	Seurat CCA
CMOT (p=100%)	1.29×10^{-6}	0.61	0.44	0.90	0.005
CMOT (p=75%)	5.73×10^{-6}	0.73	0.57	0.90	0.012
CMOT (p=50%)	5.7×10^{-4}	0.96	0.91	1.00	0.14
CMOT (p=25%)	0.06	0.99	0.99	1.00	0.74

Table A.21: **A549 gene-wise correlation** [17]. Number of genes CMOT¿m vs CMOT¡m and Wilcox–rank-sum p-values (alternative=`greater`).

Method m	# Genes (CMOT¿m)	# Genes (CMOT¡m)	P-value
MOFA+	636	547	2.19×10^{-6}
Seurat	435	748	1

Table A.22: **Pan-cancer gene expression (Pearson)** [101]. Mean and median cell-wise Pearson correlation comparing CMOT(p=25%,50%,75%,100%), Seurat, and MOFA+ (Fig. 5A).

Methods	Mean Pearson Corr.	Median Pearson Corr.
CMOT (p=100%)	0.67	0.69
CMOT (p=75%)	0.66	0.68
CMOT (p=50%)	0.62	0.63
CMOT (p=25%)	0.60	0.63
Seurat	0.63	0.65
MOFA+	0.47	0.55

Table A.23: **Pan-cancer gene expression (Spearman)** [101]. Mean and median cell-wise Spearman correlation comparing CMOT(p=25%,50%,75%,100%), Seurat, and MOFA+ (Fig. 5A).

Methods	Mean Spearman Corr.	Median Spearman Corr.
CMOT (p=100%)	0.69	0.71
CMOT (p=75%)	0.67	0.69
CMOT (p=50%)	0.65	0.68
CMOT (p=25%)	0.63	0.66
Seurat	0.64	0.67
MOFA+	0.50	0.56

Table A.24: **Pan-cancer Pearson p-values** [101]. Wilcox–rank-sum p-values comparing CMOT, Seurat, MOFA+ (Fig. 5A, Table S22).

Methods	Seurat	MOFA+
CMOT (p=100%)	7.4×10^{-4}	2.5×10^{-6}
CMOT (p=75%)	0.011	4.5×10^{-5}
CMOT (p=50%)	0.83	0.002
CMOT (p=25%)	0.98	0.017

Table A.25: **Pan-cancer Spearman p-values** [101]. Wilcox–rank-sum p-values comparing CMOT, Seurat, MOFA+ (Fig. 5A, Table S23).

Methods	Seurat	MOFA+
CMOT (p=100%)	0.006	1.5×10^{-5}
CMOT (p=75%)	0.033	8.6×10^{-5}
CMOT (p=50%)	0.50	0.002
CMOT (p=25%)	0.80	0.01

Table A.26: **Pan-cancer silhouette p-values** [101]. Wilcox–rank-sum p-values for silhouette scores (Fig. 5B).

Methods	Seurat	MOFA+	Measured RNA	Measured ATAC
CMOT (p=100%)	1.03×10^{-5}	1.6×10^{-10}	1.5×10^{-18}	5.38×10^{-18}

Table A.27: **Pan-cancer chromatin (Pearson)** [101]. Mean and median cell-wise Pearson correlation for chromatin accessibility inference comparing CMOT, Seurat, and MOFA+ (Fig. S4). All tests use alternative=`greater`.

Methods	Mean Pearson Corr.	Median Pearson Corr.
CMOT (p=100%)	0.35	0.29
CMOT (p=75%)	0.31	0.26
CMOT (p=50%)	0.30	0.23
CMOT (p=25%)	0.28	0.21
Seurat	0.41	0.37
MOFA+	-0.03	-0.03

Table A.28: **Pan-cancer chromatin (Spearman)** [101]. Wilcox–rank-sum p-values of cell-wise Pearson correlation comparing CMOT, Seurat, and MOFA+ (Fig. S4). All tests use alternative=`greater`.

Methods	Seurat	MOFA+
CMOT (p=100%)	0.18	0.73
CMOT (p=75%)	0.09	0.59
CMOT (p=50%)	0.32	0.80
CMOT (p=25%)	0.99	0.99

Table A.29: **Running times (seconds) of all methods across datasets**. Developmental human brain [149], mouse brain [26], PBMCs [54], DEX-treated A549 [17], pan-cancer gene expression [101], and chromatin inference [101].

Method	Dev. Brain	Mouse Brain	PBMCs	A549	Pan-Cancer (RNA)	Pan-Cancer (ATAC)
CMOT	22.62	19.56	84.14	2.93	0.96	0.73
Seurat	49.74	–	347.05	18.35	11.46	13.61
MOFA+	264.13	–	1953.2	272.3	226.54	164.91
TotalVI	–	–	426.36	–	–	–
Polarbear	–	964	–	–	–	–
Polarbear-coassay	–	292	–	–	–	–
scGLUE	844.12	1011	–	710.79	–	–
bindSC	951.3	2873	–	19.82	–	–
Seurat CCA	122.94	491.13	–	10.13	–	–

Supplementary Figures

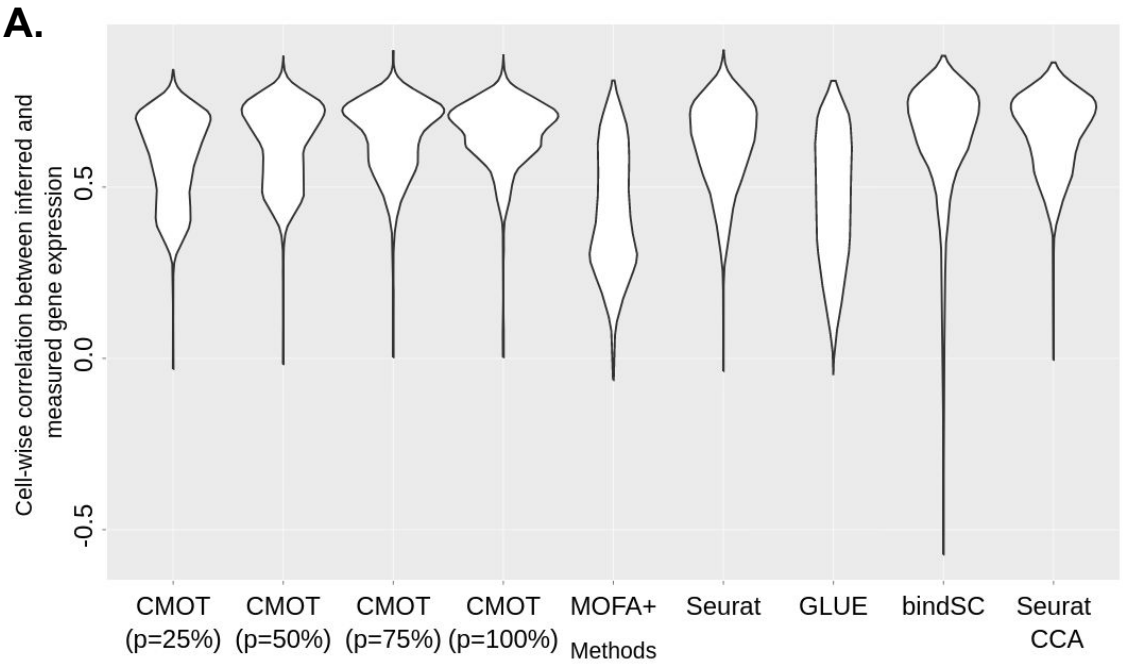

Figure A.1: **Gene expression inference from chromatin accessibility in human developing brain data [149]**. (A) Cell-wise Pearson correlation (y-axis) of inferred and measured chromatin accessibility by different methods (x-axis): CMOT (p=25%,50%,75%,100%), Seurat, MOFA+, GLUE, bindSC, Seurat CCA (Tables S1–S4).

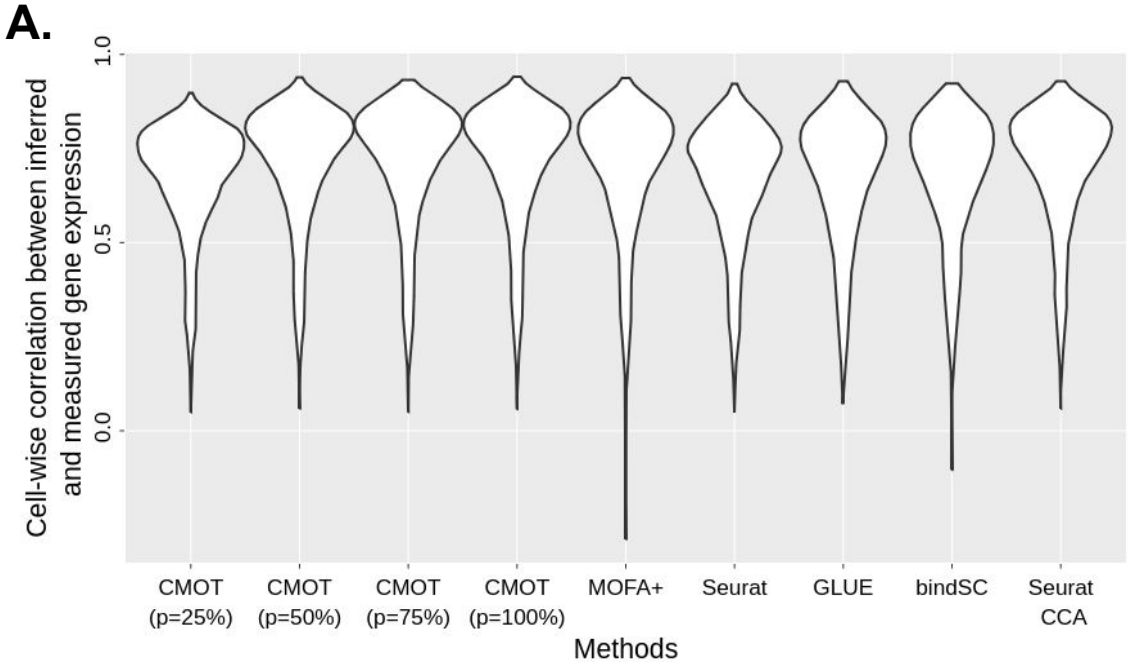

Figure A.2: **Gene expression inference from chromatin accessibility in mouse brain data [26]**. (A) Cell-wise Pearson correlation (y-axis) of inferred and measured gene expression by CMOT (p=25%,50%,75%,100%), Seurat, MOFA+, GLUE, bindSC, Seurat CCA (Tables S5–S8).

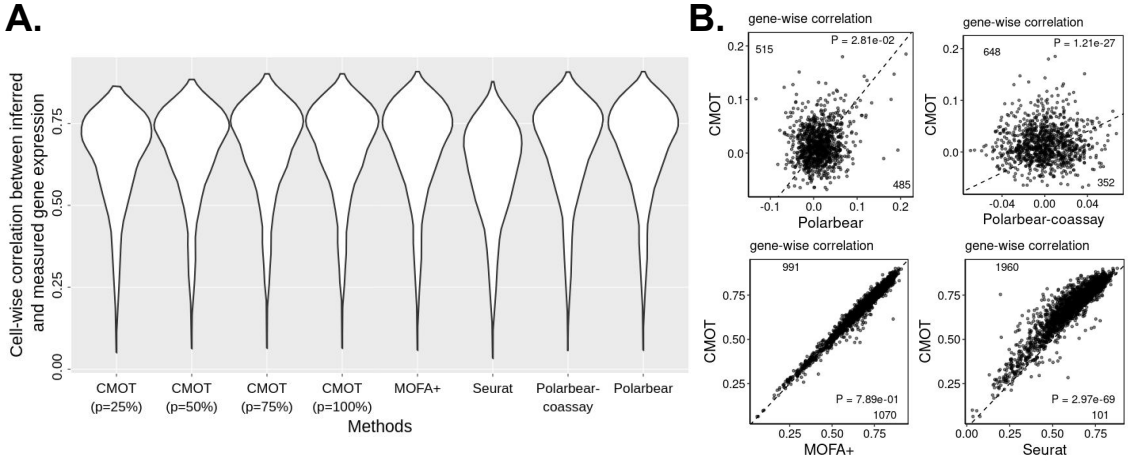

Figure A.3: **Gene expression inference from chromatin accessibility in mouse brain data [26]**. (A) Cell-wise Pearson correlation (y-axis) of inferred and measured gene expression by CMOT (p=25%,50%,75%,100%), Seurat, MOFA+, Polarbear, Polarbear-coassay (Tables S9–S10). (B) Gene-wise correlation between inferred and measured expression: CMOT (y-axis) vs. Polarbear, Polarbear-coassay, MOFA+, Seurat (x-axis). Dots are genes; numbers are gene counts above/below the dotted line.

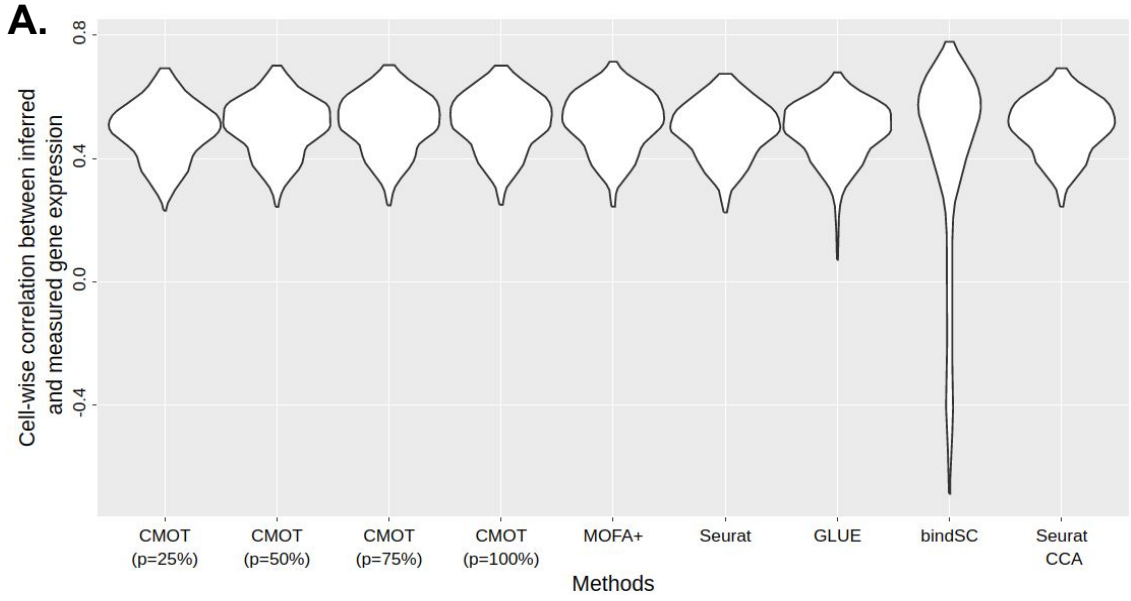

Figure A.4: **Gene expression inference from chromatin accessibility in DEX-treated A549 cells [17]**. (A) Cell-wise Pearson correlation (y-axis) of inferred vs. measured chromatin accessibility by CMOT (p=25%,50%,75%,100%), MOFA+, Seurat, GLUE, bindSC, Seurat CCA (Tables S17–S20).

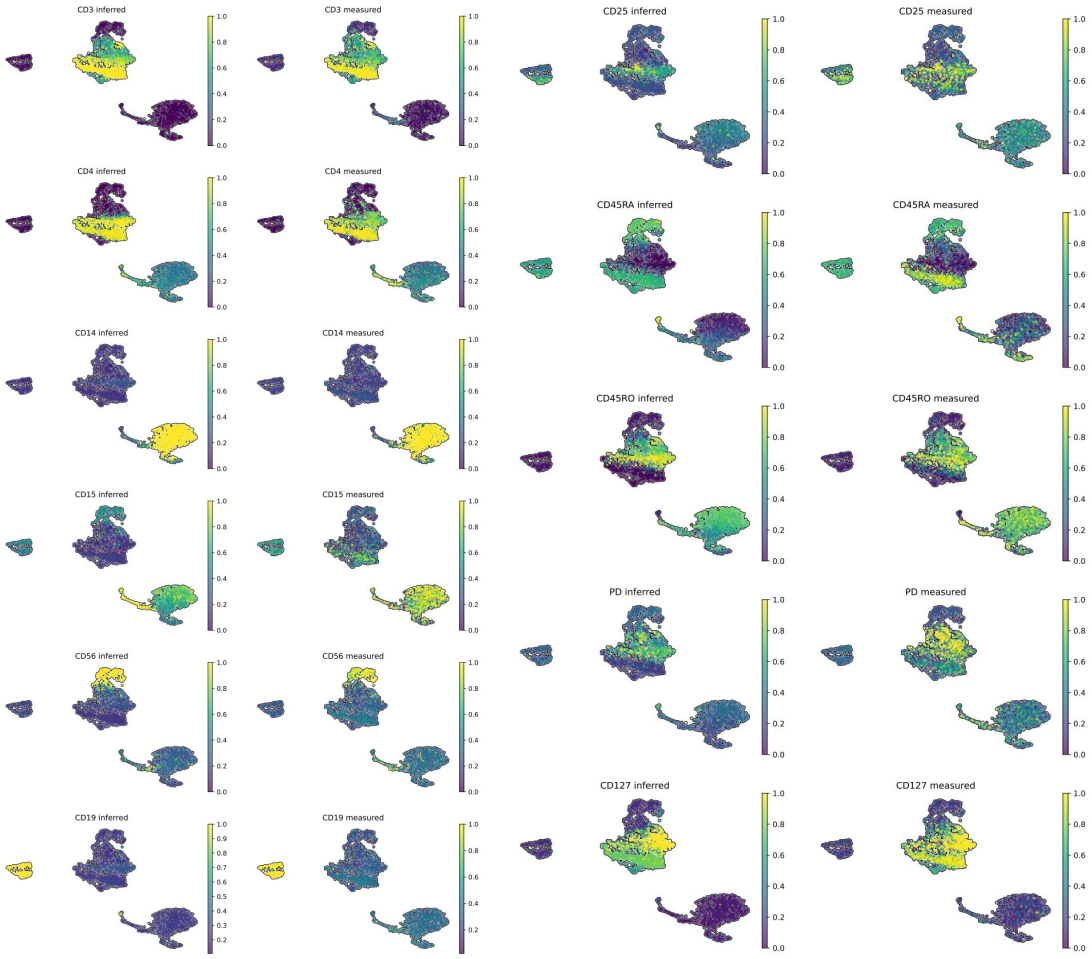

Figure A.5: **Inferring protein expression from RNA in PBMCs [54]**. Inferred vs. measured protein expression for peripheral blood mononuclear cells by CMOT.

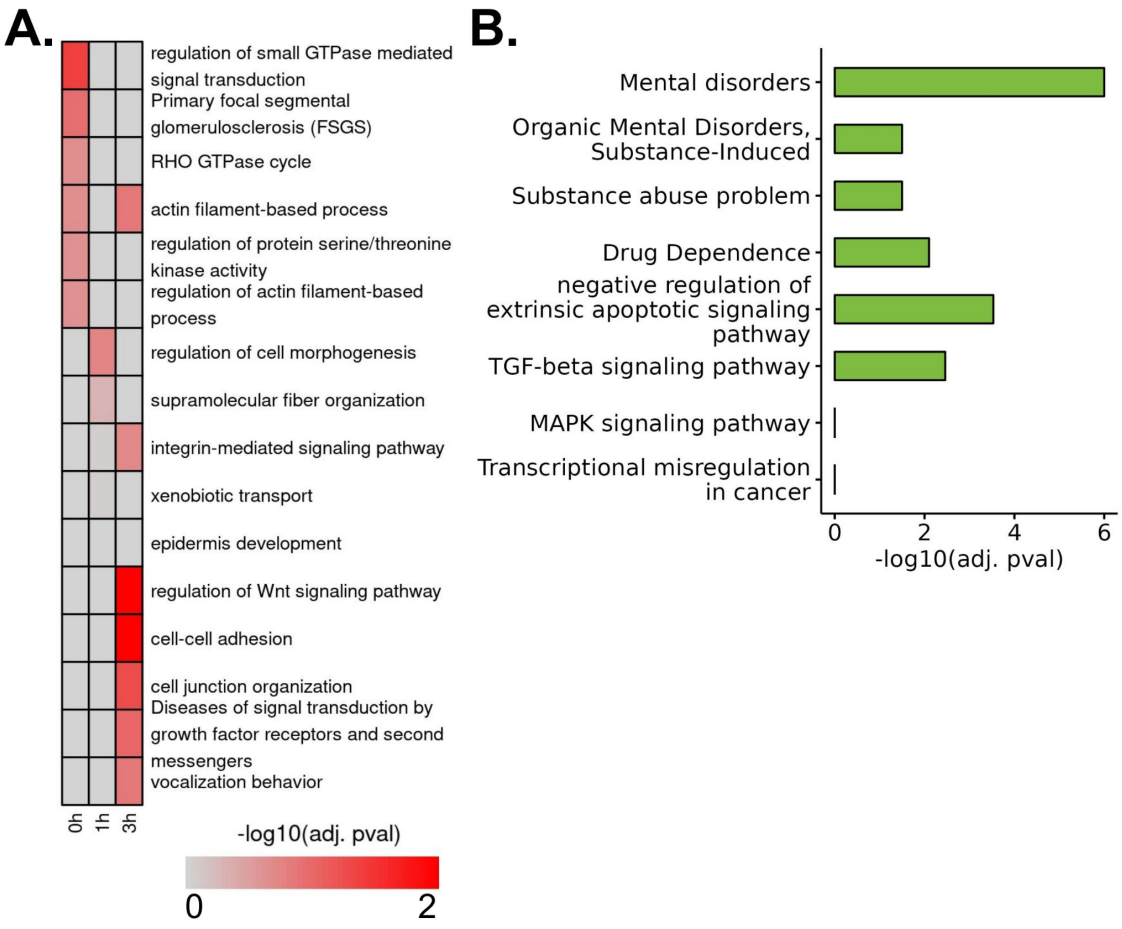

Figure A.6: **Inference of gene expression for DEX-treated A549 lung cancer cells [17].** (A) Heatmap of top 100 predictive genes' enriched terms ranked by $-\log_{10}$(adj. p-value). (B) Enriched terms for MOFA+ inferred expression using 748 genes with higher gene-wise Pearson correlation than CMOT's 435 genes (Fig. 4B,D).

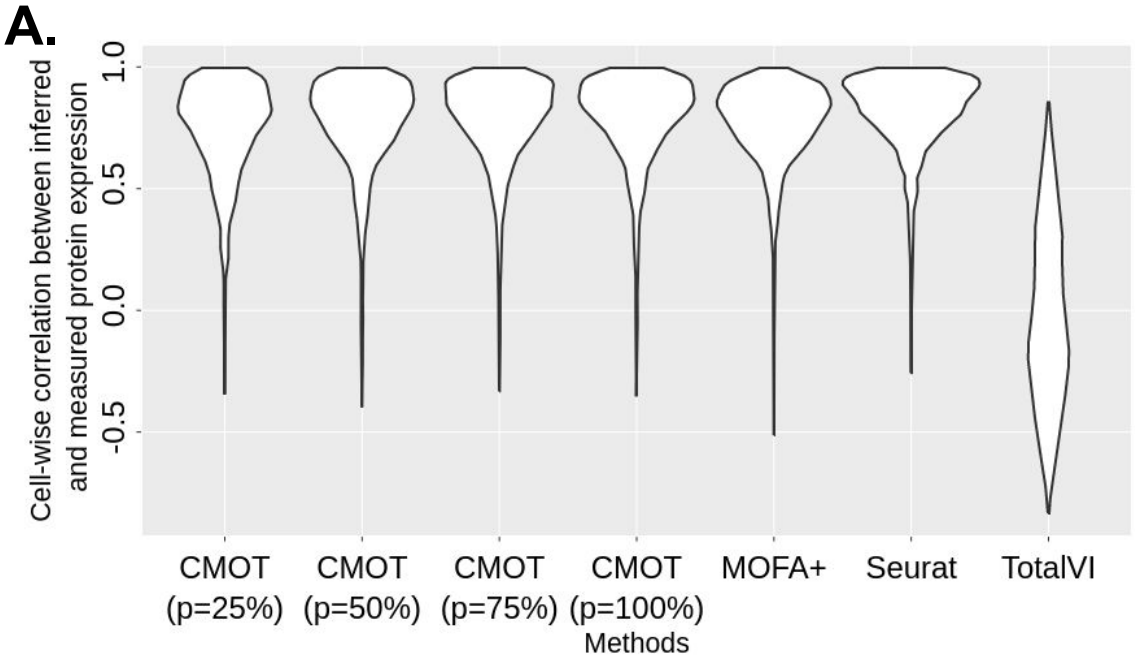

Figure A.7: **Cross-modality inference between protein and gene expression in PBMC10k [54]**. Cell-wise Pearson correlation (y-axis) of inferred vs. measured protein expression: CMOT (p=25%,50%,75%,100%), Seurat, MOFA+, TotalVI (Tables S14–S16).

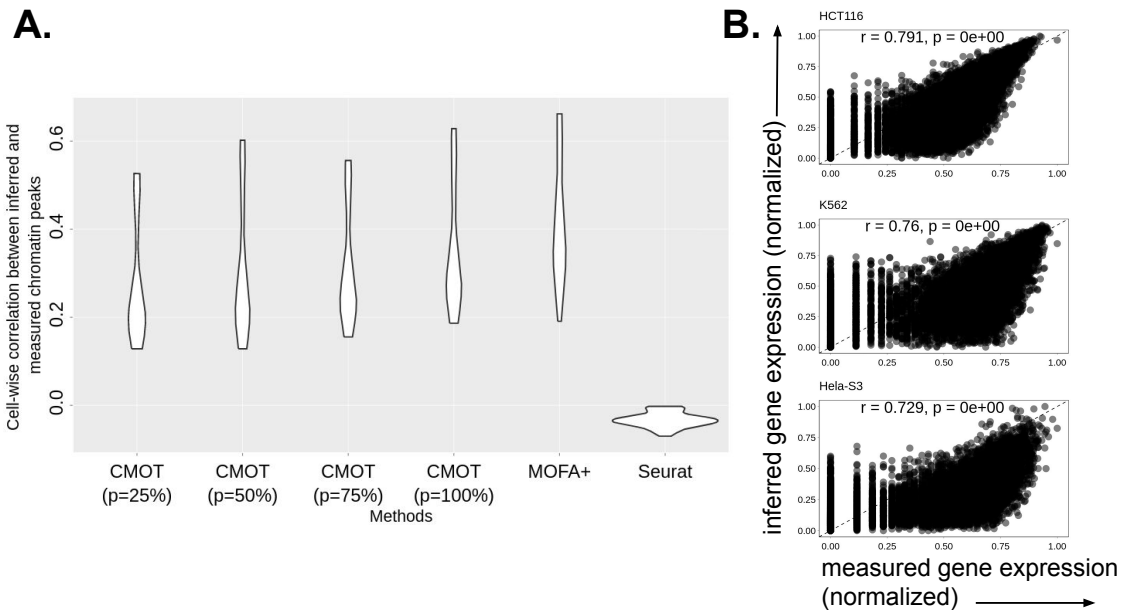

Figure A.8: **Cross-modality inference between gene expression and chromatin accessibility in pan-cancer cells [101]**. (A) Cell-wise Pearson correlation of inferred vs. measured chromatin accessibility by CMOT (p=25%,50%,75%,100%), Seurat, MOFA+ (Tables S27–S28). (B) Measured (x-axis) vs. inferred normalized expression (y-axis) of genes (dots) for three example cells; r is Pearson's r, p is the correlation p-value.

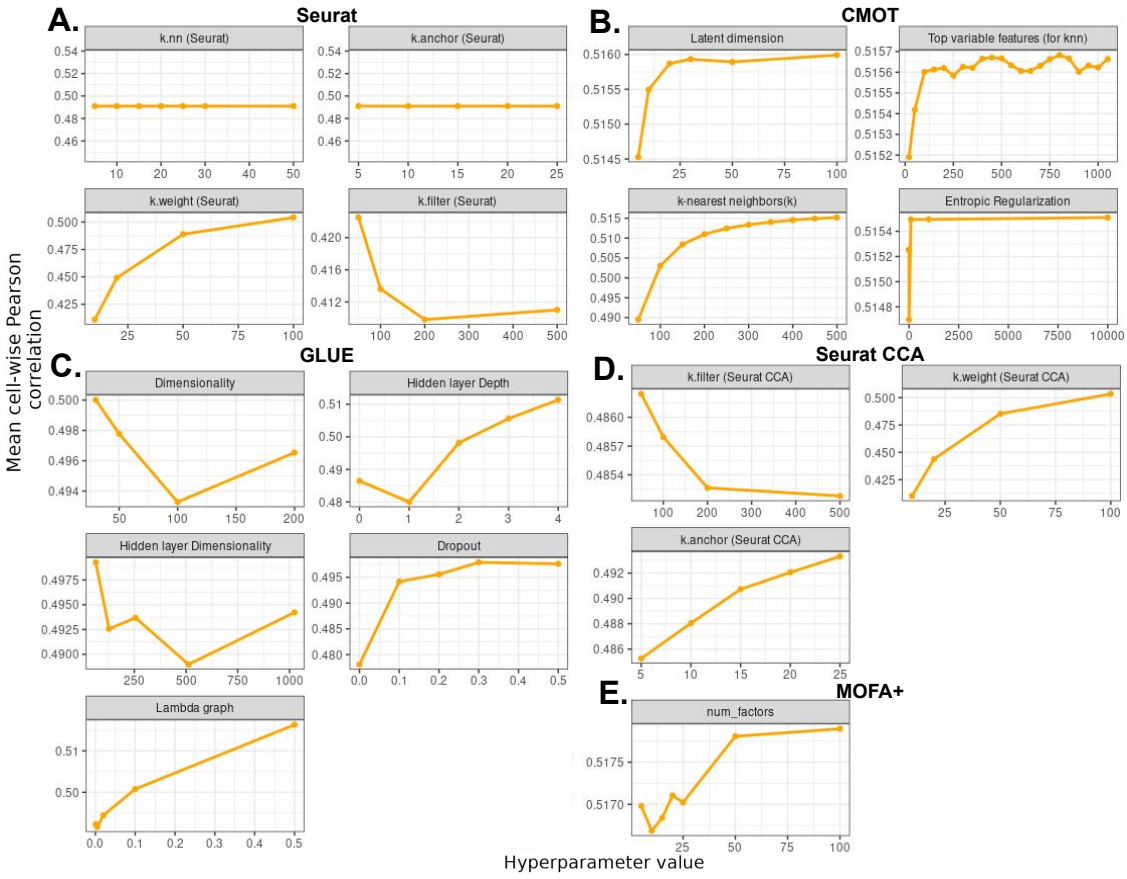

Figure A.9: **Hyperparameter sensitivity on DEX-treated A549 [17]**. (A) Seurat: k.nn, k.anchor, k.weight, k.filter. (B) CMOT: latent dim, # features, k.neighbors, entropic regularization. (C) GLUE: embedding dim, hidden depth, hidden dim, dropout, λ_{graph}. (D) Seurat CCA: k.anchor, k.weight, k.filter. (E) MOFA+: num_factors.

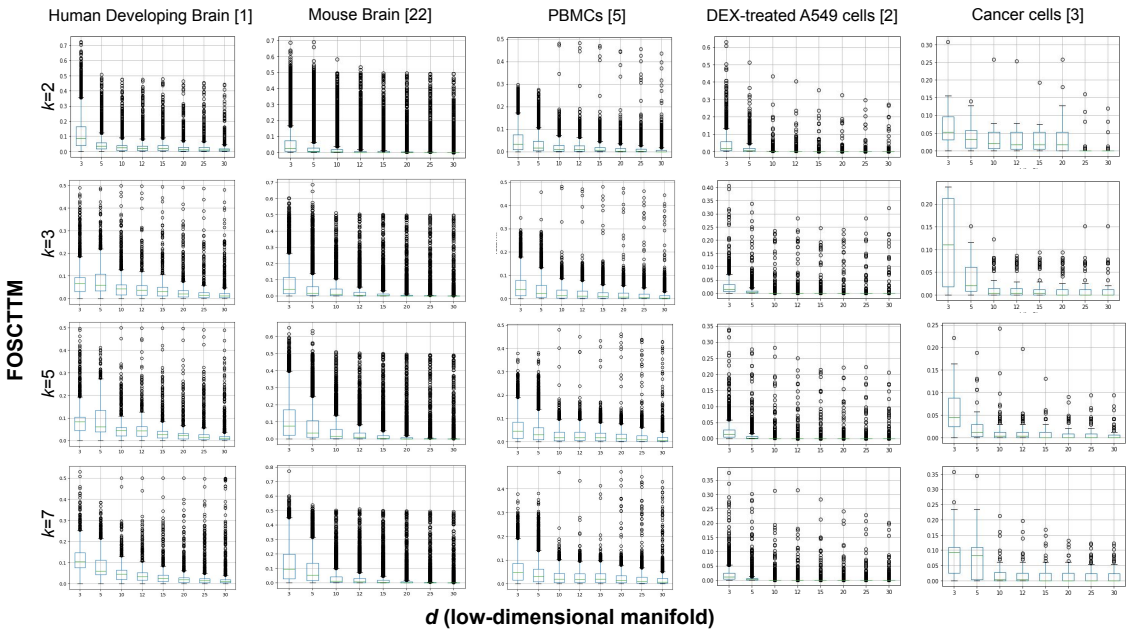

Figure A.10: **FOSCTTM scores across NMA methods**. Boxplots of pairwise cell Mean FOSCTTM score (y-axis) vs. latent dimension d (x-axis), for varying k-nearest neighbors (rows) and datasets (columns).

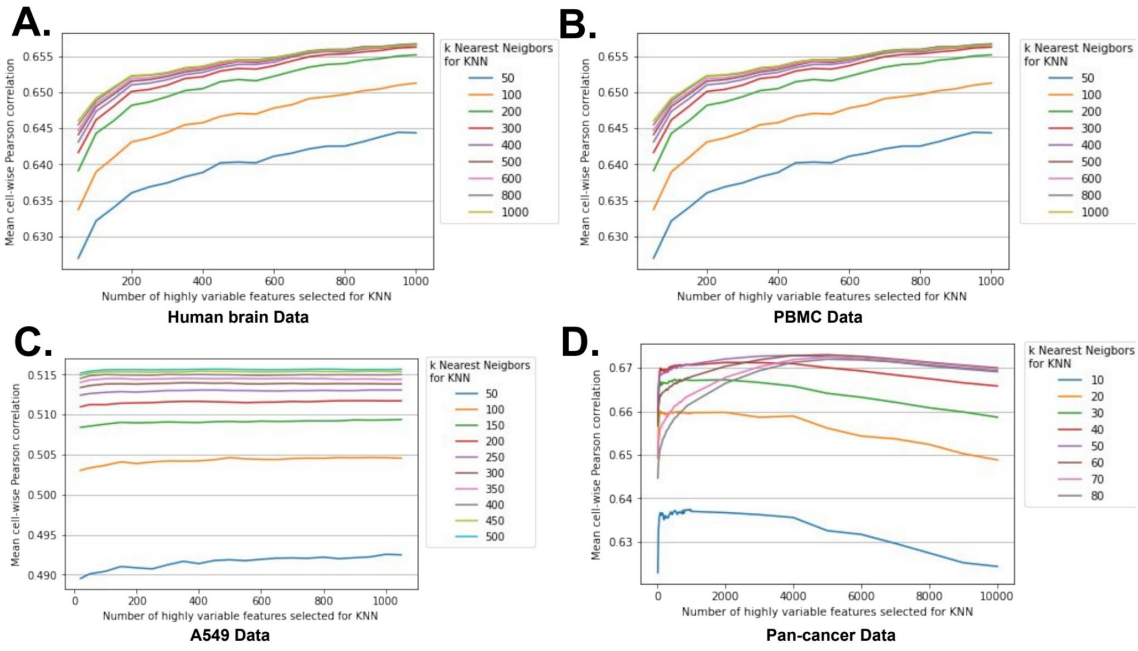

Figure A.11: **Mean Pearson correlation across variable genes**. (A) Human brain [149]. (B) PBMC [54]. (C) A549 [17]. (D) Pan-cancer [101]. Each panel: mean cell-wise Pearson correlation vs. # top highly variable genes.

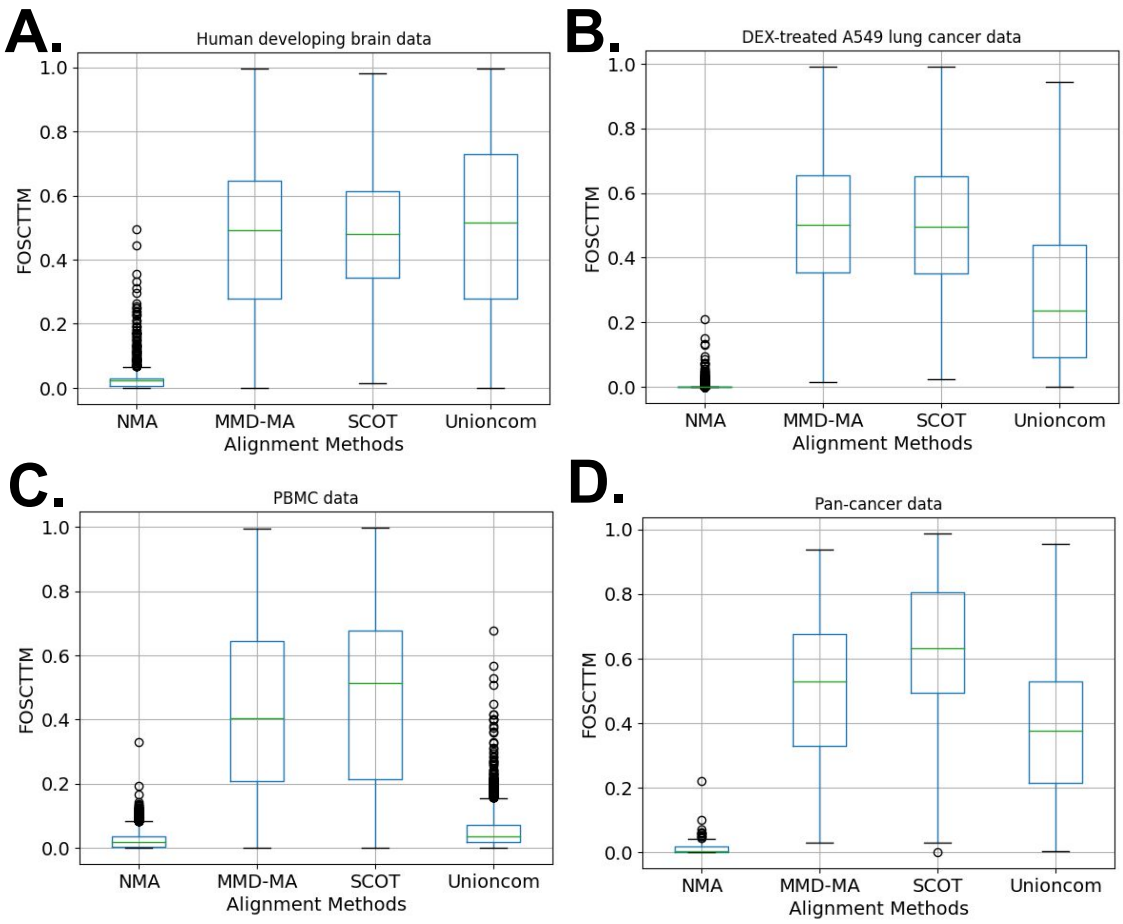

Figure A.12: **FOSCTTM for alignment methods**. Boxplots of FOSCTTM for NMA, MMD-MA, SCOT, Unioncom on four datasets with optimal d (human brain d=20, PBMC d=15, A549 d=10, pan-cancer d=10).

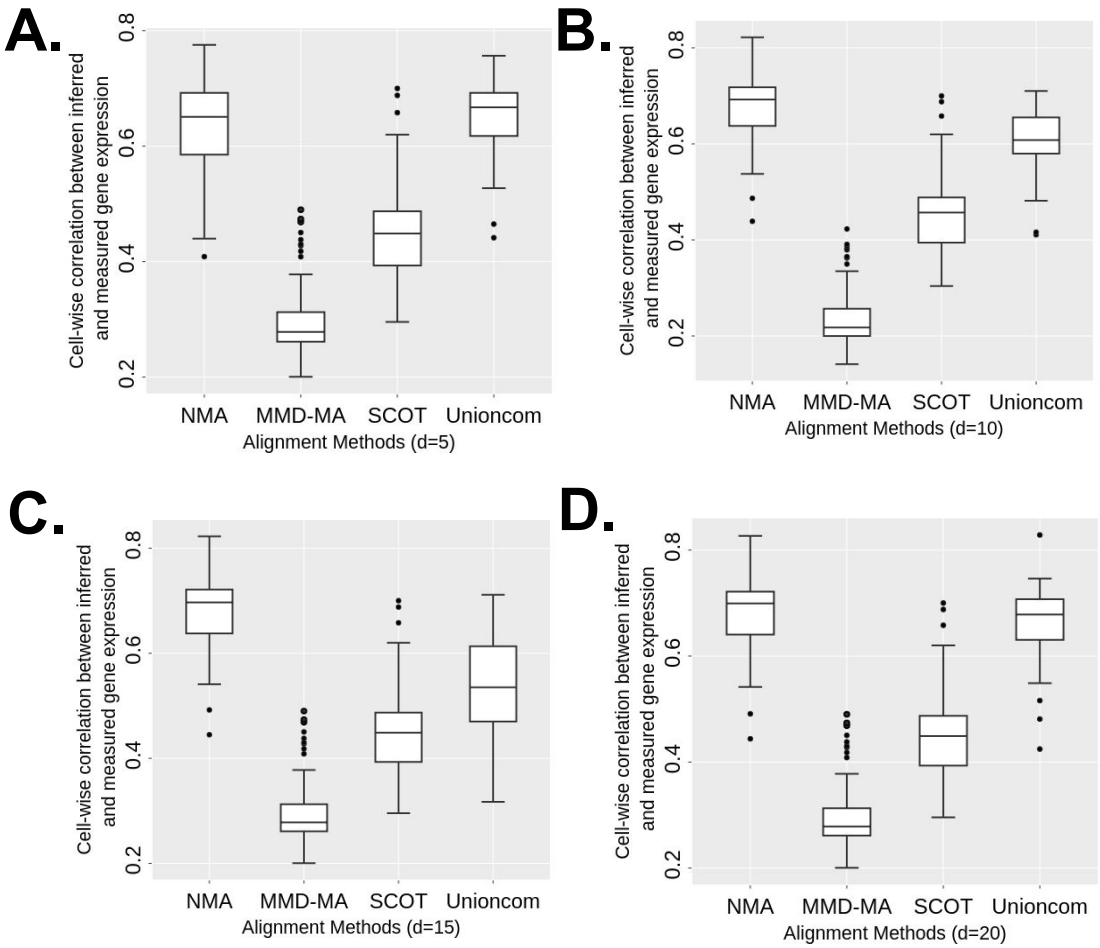

Figure A.13: **Benchmarking alignment in pan-cancer [101]**. Mean cell-wise Pearson correlation vs. alignment method: (A) d=5, (B) d=10, (C) d=15, (D) d=20.

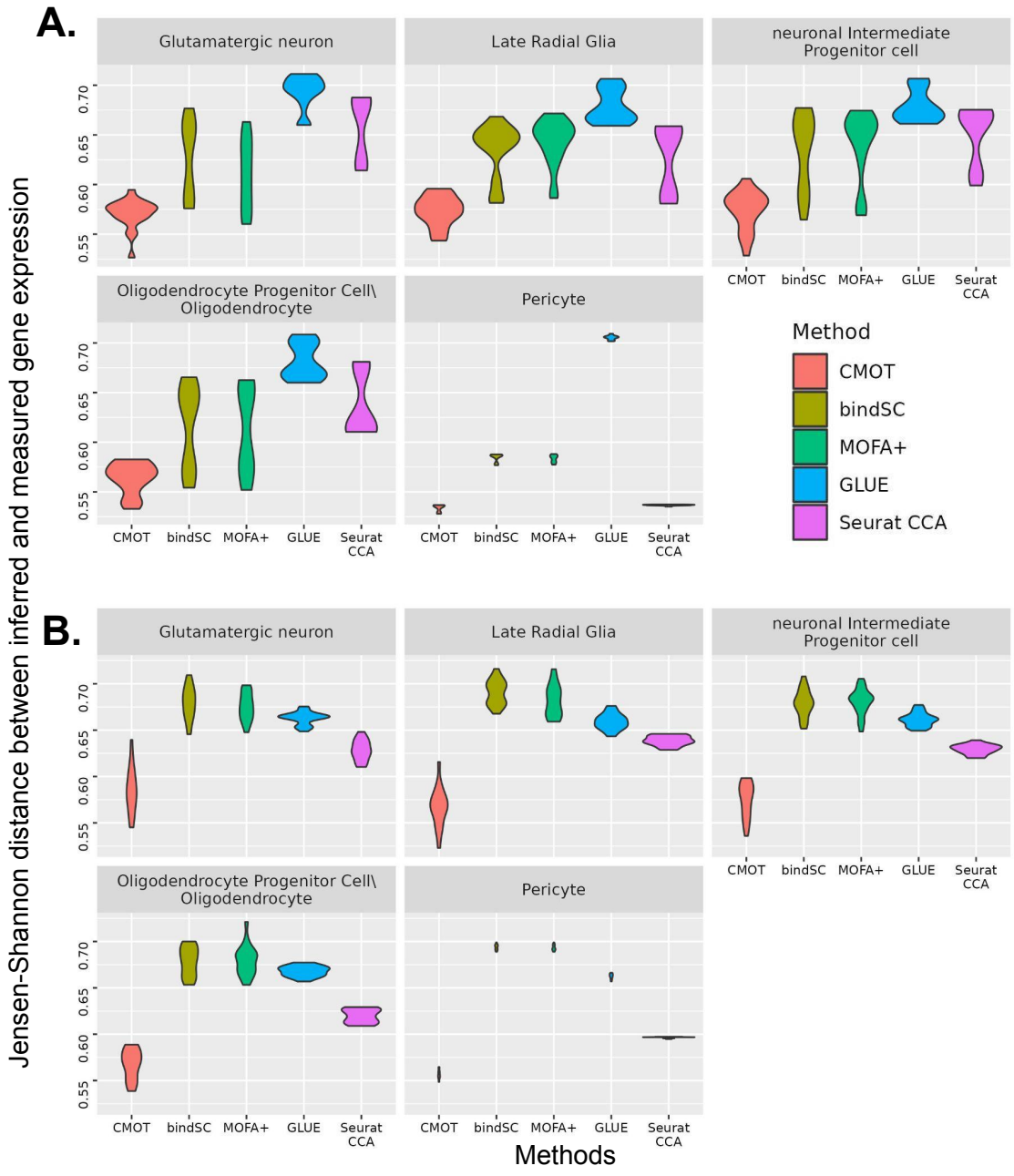

Figure A.14: **Jensen–Shannon distance in human developing brain data [149]**. (A) 1,000 HVGs: distance between measured vs. inferred by CMOT, bindSC, MOFA+, GLUE, Seurat CCA. (B) 2,000 HVGs: same comparison across five major cell types.

Supplementary Methods

Datasets preprocessing and feature selection

Mouse brain: The adult mouse brain dataset [26] was generated by SNARE-seq, containing jointly profiled gene expression and open chromatin regions for 10k cells. However, we used the previously processed dataset by Cao et al. [17] (published with GLUE), including gene expression with 28,930 genes and open chromatin peaks with 241757 peaks for 9190 cells from the adult mouse brain. However, we reduced the size of the dataset by picking 2000 highly variable genes and peaks. The resulting data includes gene expression and open chromatin peaks of 9190 cells for 2000 genes and peaks. To compare CMOT with Polarbear we used their data with 10k cells. However, we reduced the number of genes and peaks. We filtered out peaks and genes that occurred in less than 3 cells. For the binary scATAC data, we picked the top 1000 peaks that were expressed in the largest number of cells. For scRNA, we performed normalization and variance stabilization using SCTransform [61] and picked the top 1000 most variable genes. The resulting data includes gene expression and chromatin regions of 10,839 cells for 1000 genes and regions respectively.

Peripheral Blood Mononuclear Cells: The PBMC10k dataset [54] contains 6855 cells, containing genes and proteins from the same cells. For scRNAseq, we performed normalization and variance stabilization using SCTransform, and used the 2960 highly variable genes. For protein expression, we performed centered log-ratio (CLR) normalization using Seurat's functions. The resulting dataset includes 6855 cells with 2960 genes and 14 proteins.

Inferring protein expression from gene expression in Peripheral Blood Mononuclear Cells

We applied CMOT to infer protein expression from gene expression of peripheral blood mononuclear cells (PBMCs) using emerging CITE-seq data [54]. We randomly split the PBMC10k cells into 80% training for cross-validation and 20% testing set for evaluation. We trained CMOT with parameters: K=5, d= 15, λ=1e02, $\eta = 1$, k=100, and used the top 200 highly variable genes in the training data to find the k nearest neighbors. We induced cell labels by identifying two clusters using gene expression for the label regularization in optimal transport. As shown in Fig. S4, CMOT achieves a median correlation r of 0.91 for p=100% outperforming MOFA+ (median r=0.89, Wilcoxon $p - values < 1.22e - 05$) and TotalVI (median r=-0.08, Wilcoxon p-values=0) and performs comparably with Seurat (median r=0.92, Wilcoxon $p-values < 0.99$). Also, for p=50%,75%, CMOT reports a higher correlation of 0.9 for both, outperforming both MOFA+ (Wilcoxon $p-values < 1.4e-03$) and TotalVI (Wilcoxon p-values=0) (Tables S14-S16).

Benchmark against state-of-art for cross-modality inference

We benchmarked CMOT against state-of-art methods: Seurat [139], Seurat CCA [66], MOFA+ [9], bindSC [43], GLUE [20], and, Polarbear [167]. We show bench-

marking on 3 datasets: (1) mouse brain data [26] (GLUE's application), (2) DEX-treated A549 data [17] (common application between CMOT and bindSC), and, (3) human developing brain [149](our application),

For the mouse brain dataset [26], which is an application of GLUE [21], we used their preprocessed dataset to benchmark all methods. We split this dataset into 80% train and 20% test, and benchmarked all methods on default parameters. For CMOT, we used the following parameters: K=5, d= 20, λ=1e03, η=1e-2, k=400, and used the top 10 features of scATAC-seq training data to find the k nearest neighbors. We induced our cell labels by hierarchical clustering of the cells in the training set and identified two clusters to regularize the optimal transport in CMOT. As shown in Fig S2, CMOT outperforms state-of-art methods with a median correlation of 0.76 compared to Seurat (median correlation=0.7, Wilcoxon $p-value < 3.4e-40$), MOFA+(median correlation=0.74, Wilcoxon $p-value < 1.15e-05$), GLUE(median correlation=0.72, Wilcoxon p-value¡$5.6e-25$), bindSC (median correlation=0.72, Wilcoxon p-value¡8e-19), and, Seurat CCA (median correlation=0.75, Wilcoxon p-value¡1.9e-04). Even for partial correspondences, CMOT reports a higher performance (median r=0.76 for p=75%, median r=0.76 for p=50%) continuing to outperform other methods: Seurat (Wilcoxon p-value¡1.25e-39), GLUE (Wilcoxon $p-value < 1.63e-24$), MOFA+ (Wilcoxon $p-value < 1.67e-05$), bindSC (Wilcoxon $p-value < 2.07e-158$), Seurat CCA (Wilcoxon $p-value < 2.8e-4$). Even for correspondence as low as 25%, CMOT performs comparably with median r=0.71, still outperforming Seurat (Wilcoxon p-value¡0.07) (see Additional File 1: Fig S2, Supplemental Methods, Supplemental Table S5-S8). We also compared CMOT with Polarbear on their mouse brain data for fair comparison using top 1000 features. We found that CMOT has significantly higher gene-wise correlations than Polarbear (515 genes versus 485 genes, Wilcoxon $p-value < 2.81e-02$) and Polarbear co-assay (648 genes versus 352 genes, Wilcoxon $p-value < 1.2e-27$), and, Seurat (1960 genes versus 101 genes, Wilcoxon $p-value < 2.97e-69$) (Fig. S3, Tables S9-S10).

For the DEX-treated A549 dataset [17], which is a common application of CMOT and bindSC, we also split the dataset into 80% train and 20% test, similar to Fig 4. To run bindSC, we used the parameters suggested in their dataset tutorial [43]. We benchmarked other methods on tuned parameters on 80% of training data (Fig S9). After tuning, we used the following parameters for benchmarking: A) Seurat: k.nn=20, k.anchor=5, k.weight=100, k.filter=50; B) CMOT: latent dimension=25, top variable features=20, k-nearest neighbors=500, entropic regularization=100; C) GLUE: Dimensionality=30, Hidden layer depth=4, Hidden layer dimensionality=64, Dropout=0.3, Lambda graph=0.5; D) Seurat CCA: k.filter=50, k.weight=100, k.anchor=25; E) MOFA+: num_factors=100. As shown in Additional File 1: supplemental Fig S7, CMOT outperforms state-of-art with a median correlation r of 0.52, compared to Seurat (median r=0.5, Wilcoxon $p-value < 21.27e-05$), GLUE(median r=0.5, Wilcoxon $p-value < 4.7e-06$), bindSC (median r=0.51, Wilcoxon $p-value < 0.016$), Seurat CCA (median r=0.51, Wilcoxon

$p - value < 0.016$) and performs comparably to MOFA+(median r=0.52, Wilcoxon $p - value < 0.64$). Even for partial correspondences, CMOT reports high and consistent performance (median r=0.52 for p=75%, median r=0.51 for p=50%) outperforming Seurat (Wilcoxon $p - value < 3.9e - 05$), GLUE (Wilcoxon $p - value < 1.6e - 05$), bindSC (Wilcoxon $p - value < 0.025$), and, Seurat CCA (Wilcoxon $p - value < 0.029$). (see Fig S7, Tables S17-S20).

To benchmark the human developing brain data [149], we split the dataset into 80% train and 20% as in Fig. 1. We trained all methods on default parameters and for CMOT, we used the same parameters as in Fig. 2. As shown Figure A.1, CMOT outperforms or performs comparably with state-of-arts with a median correlation of 0.67, compared to Seurat (median correlation=0.64, Wilcoxon $p - value < 1.23e - 14$), MOFA+(median correlation=0.41, Wilcoxon p-value=0), GLUE(median correlation=0.47, Wilcoxon $p - value < 1.4e - 236$), bindSC (median correlation=0.68, Wilcoxon $p - value < 0.97$), and Seurat CCA (median correlation=0.68, Wilcoxon $p - value < 0.98$). Also, for p¡100%, CMOT continues to report consistent performance. For example, CMOT has significantly higher performances (median r=0.65 for p=75%, and r=0.63 for p=50%) than MOFA+ (Wilcoxon $p - value < 2.8e - 294$), GLUE (Wilcoxon $p - value < 1.3e - 216$), and Seurat (Wilcoxon $p - value < 3.43e - 10$). Also, with low correspondence such as p=25%, CMOT's performance (r=0.61) is still significantly higher than MOFA+ (Wilcoxon $p - value < 1.65e - 157$) and GLUE (Wilcoxon $p - value < 2.4e - 86$) (see Fig S1, Tables S1-S4).

Benchmarking on scRNA-seq and scATAC-seq datasets

We next benchmarked CMOT's performance on single profiled scRNA-seq and scATAC-seq datasets from the developing human brain (week 21) [149]. We picked the top five major cell types common to both datasets: Glutamatergic neuron, Late Radial Glia, neuronal Intermediate Progenitor cells, Oligodendrocyte, and Pericyte. This resulted in 7420 cells from scATAC-seq and 6542 cells from scRNA-seq. We preprocessed each dataset similar to previous datasets, i.e. for scRNA-seq, we performed normalization and variance stabilization using SCTransform, and for scATAC-seq, we normalized the peaks using term frequency-inverse document frequency (TF-IDF) transformation using RunTFIDF [45]. Also, we conducted two evaluations by selecting 1000 and 2000 highly variable features. Finally, we split each profile into 80% training and 20% testing sets and evaluated CMOT against bindSC, MOFA+, GLUE, and Seurat CCA (Fig. S14). We ran all methods on default or recommended parameter settings. For CMOT, we used the same parameters as those used for the first application on developing humain brain. It is important that the inferred expressions must preserve the cell type distributions, hence we computed the Jensen-Shannon distance between cells from each cell type. We found that CMOT outperformed the state-of-art methods. Jensen-Shannon distance (JS distance): The Jensen-Shannon distance measures the similarity between two probability distributions, and is the square root of Jensen-Shannon divergence (JSD).

JSD is the average KL divergence between two distributions and their average (also the symmetric version of the KL divergence). KL divergence measures the difference between two probability distributions and increases proportionally. The smaller the JS distance, the closer are two probability distributions.

[**?**] We benchmarked CMOT's performance across different alignment methods including Nonlinear Manifold Alignment (NMA) [75], Maximum mean discrepancy-based manifold alignment (MMD-MA) [100], Single-cell alignment with optimal transport (SCOT) [39], and Unioncom [18] (Figures A.12,-A.13) D) on the pan-cancer data [101]. We experimented with a different number of latent dimensions d=5,10,15,20 while using the default parameters for all methods. We evaluated CMOT's performance on the pan-cancer data across all dimensions and found that using NMA gives the best inference results (A.13). Additionally, we also evaluate the performance of alignment methods using the FOSCTTM score across all datasets (A.12) for latent dimensions reported in the Results. We see that NMA significantly outperforms all other alignments.

Outlier cell detection in target modality

We tested our outlier detection mechanism in the DEX-treated A549 dataset [17], where we randomly replaced 25% of cells in the test dataset with noisy cell samples. We generated the noisy cell samples by generating random real numbers within the interval of minimum and maximum values of the normalized chromatin expression. The IF mechanism can successfully identify the noisy cells, with an AUC=0.99.

A.2 ARTEMIS: Supplement

Algorithm S1 ARTEMIS

Input: $\{X_t\}_{t\in\{0,1,..T\}}, t \in \{0, 1, ..T\}$, prior kill rate k,
pretrain epochs E_1, number of iterations E_2, joint train epochs E_3,
#SDEs to sample E_4

Initialize: $q_\varphi, p_\phi, Q_\theta, \widehat{Q}_{\widehat{\theta}}, K_\omega$

1: Pre-train VAE:
2: **for** i=1 to E_1 **do**
3: $q_\varphi(X_t||t) = (\mu_t, \sigma_t)$, $Z_t \sim \mathcal{N}(\mu_t, \sigma_t^2)$, $\widehat{X}_t = p_\phi(Z_t)$
4: Update φ, ϕ using ∇L_{vae} (Eq. 2)
5: **end for**
6: Jointly train VAE and uDSB:
7: **for** j=1 to E_2 **do**
8: **for** k=1 to E_3 **do**
9: **for** l=1 to E_4 **do**
10: $(\overleftarrow{Z}_t, A_t) \leftarrow$ sample B-SDE$(Q_\theta, \widehat{Q}_{\widehat{\theta}}, K_\omega)$ (Eq. 9b)
11: Update θ using $\nabla_\theta L_{div,\theta}$ (Eq. 12b)
12: sample F-SDE $\overleftarrow{Z}_t \sim \rho_0$ (Eq. 9a)
13: update $\theta, (\varphi, \phi)$ using $\nabla_{\theta,\varphi,\phi} L_{joint}$ (Eq. 13)
14: **end for**
15: **for** l=1 to E_4 **do**
16: $(\overrightarrow{Z}_t, A_t) \leftarrow$ sample F-SDE$(Q_\theta, \widehat{Q}_{\widehat{\theta}}, K_\omega)$(Eq. 9a)
17: update $\widehat{\theta}, \omega$ using $\nabla_{\widehat{\theta}} L_{div,\widehat{\theta}}$ (Eq. 12a),$\nabla_\omega L_\omega$ (Eq. eq:10), respectively
18: **end for**
19: **end for**
20: **end for**
21: **Outputs:**
$\varphi, \phi, \theta, \widehat{\theta}, \omega$

S1 Details of model implementation, training, and testing

ARTEMIS is implemented in JAX [16]. Training begins with VAE pre-training on single-cell gene expression data from observed timepoints. The ADAM optimizer with gradient clipping (threshold=1) and initial learning rate as 0.0001 was used and the VAE was trained for E_1 epochs with a default batch size of 64 (Algorithm S1, Steps 1-5).

Subsequently, the uDSB and VAE are jointly trained to learn a smooth latent space. The uDSB model comprises three networks, a.) forward drift Q_θ, b.) backward drift $\widehat{Q}_{\widehat{\theta}}$, and c.) kill rate K_ω.

The architecture of Q_θ and $\widehat{Q}_{\widehat{\theta}}$ includes:

i) `x_encoder`: 3-layer MLP taking Z_t as input, ii) `t_encoder`: 3-layer MLP taking the sinusoidal embedding of time t as input, and, iii) `decoder`: 2-layer MLP decoder combining ouputs from `x_encoder` and `t_encoder` to output optimal drift values (Q_θ,$\widehat{Q}_{\widehat{\theta}}$).

Algorithm S2 SDE Sampling Procedure

Input: drift f, diffusion coefficient ε, killing rate r‘, initial cells n_0, final cells n_T

procedure F-SDE($Q_\theta, \widehat{Q}_{\widehat{\theta}}, K_\omega$)
  sample $\overrightarrow{Z}_0 \sim \rho_0$
  $A_0 \leftarrow 1$
  for step i=1 to T **do**
    $\overrightarrow{Z}_i = \overrightarrow{Z}_{i-1} + \Delta\overrightarrow{Z}_i$
    if $A_{i-1} = 1$ **then**
      $D \sim$ Bernoulli$(1 - k\text{‘}(i)\Delta t)$
      $A_{i+1} \leftarrow D$
    end if
  end for
end procedure
procedure B-SDE($Q_\theta, \widehat{Q}_{\widehat{\theta}}, K_\omega$)
  sample $\overleftarrow{Z}_0 \sim \rho_T$
  $A_T \sim$ Bernoulli$(\min(1, \frac{n_T}{n_0}))$
  for step i=T-1 to 0 **do**
    $\overleftarrow{Z}_i = \overleftarrow{Z}_{i-1} - \Delta\overleftarrow{Z}_i$
    if $A_{i+1} = 0$ **then**
      $D \sim$ Bernoulli$(k\text{‘}(i)\Delta t)$
      $A_i \leftarrow D$
    end if
  end for
end procedure
Outputs:
  $(Z_t, A_t)_t$

The K_ω network is a 2-layer MLP taking a sinusoidal embedding of time t and outputting a kill rate at time t. `SiLU` activation was used for Q_θ and $\widehat{Q}_{\widehat{\theta}}$; `Leaky-ReLU` for K_ω. All networks used 16-dimensional time embeddings. The ADAM optimizer with gradient clipping was used, with initial learning rates of 0.0001 for all networks (needs to be tuned according to the dataset).

uDSB training is performed within the latent space from the pre-trained VAE. The uDSB training used a batch size of 512 over E_2 iterations, each iteration comprising E_3 epochs of forward and backward training. The uDSB model was trained using the Iterative Proportional Fitting algorithm (IPF), which iteratively solves the schrödinger bridge problem through forward and backward SDEs[48, 89, 129]. During each epoch, the forward and backward drifts, as well as the VAE parameters, are updated (Algorithm S1, Steps 6-20). Specifically, in each epoch, 10 SDEs were sampled to optimize forward and backward drifts each. The process is described below:

1. Forward optimization: This step minimizes the KL divergence with a fixed terminal condition (e.g., $\overleftarrow{Z}_t \sim p_T$). A backward SDE is sampled (Eq. 9b), and the divergence loss is calculated between the SDE predicted by the forward drift and the sampled backward SDE. The VAE params are optimized concurrently using the L_{joint} loss (Eq. 13, Algorithm S1, Steps 9-14))

2. Backward optimization: This step minimizes the KL divergence with a fixed initial condition (e.g., $\overrightarrow{Z}_t \sim p_0$). A forward SDE is sampled (Eq. 9a) and the divergence loss is calculated between the SDE predicted by the backward drift and the sampled forward SDE (Algorithm S1, Steps 15-18)).

The number of discretization steps was set to 100 for the inteval [0,T], so Δt=0.01.

For trajectory inference, gene expression profiles at $t = 0$ are projected into the VAE latent space. Forward SDE sampling (Eq. 9a) with the learned drift Q_θ generates continuous latent variables, decoded back into the gene expression space to reconstruct cellular trajectories.

Prediction performance on held-out timepoints was evaluated by averaging the 2-Wasserstein distance between predicted and ground-truth gene expression profiles, computed over five forward trajectory samples. The `OTT` library [35] was used to compute distances.

We performed all training and benchmakring on Linux Ubuntu machine with 256 BG RAM and NVIDIA RTX A6000 GPU with 48 GB RAM. We have reported the runtime and scalability for different training sizes (#cells) on the zebrafish dataset Figure A.21. The code has also been tested on Linux Ubuntu machine with only CPU.

S2 Baseline Methods

We compare ARTEMIS's performance with the following baseline methods:

- PRESCIENT [163]: PRESCIENT (Potential eneRgy undErlying Single Cell gradIENTs) is a generative modeling framework designed to learn differentiation landscapes from time-series scRNA-seq data. It models how cells evolve stochastically and in physical time, using a diffusion-based approach to recover a global potential function. To handle large scRNA-seq datasets, PRESCIENT models are fit on PCA projections of scaled gene expression data. The potential function is parameterized by a neural network, to allow flexible and complex landscape modeling. PRESCIENT allows for using prior knowledge of cell growth in the modeling. However, such information is not always available and we included evaluations with growth rates information when available. We used Python codes on Github (`https://github.com/gifford-lab/prescient-analysis`) to run PRESCIENT.

- MIOFlow [77]: MIOFlow (Manifold Interpolating Optimal-Transport Flow) is a computational method for modeling stochastic, continuous population dynamics from snapshots of time-series data. It combines dynamic models, manifold learning, and optimal transport techniques to interpolate between static population snapshots. Using neural ordinary differential equations (Neural ODEs) and a geodesic autoencoder (GAE), MIOFlow ensures the flow aligns with the data's manifold geometry. By operating in the autoencoder's and penalizing transport with Wasserstein distance, complex diffusion processes in

cellular dynamics. We use Python codes on GitHub (`https://github.com/KrishnaswamyLab/MIOFlow`) to run MIOFlow

- scNODE [166]: scNODE (single-cell Neural Ordinary Differential Equation) is a deep learning model that predicts and simulates single-cell gene expression at unobserved timepoints in temporal scRNA-seq data. It combines a variational autoencoder (VAE) to encode gene expression into a low-dimensional latent space with neural ordinary differential equations (ODEs) to model the temporal evolution of cells within this space. A dynamic regularization term aligns the latent dynamics with temporal data, reducing information loss between discrete timepoints and improving predictions. We use the Python codes on Github (`https://github.com/rsinghlab/scNODE`) to run scNODE.
- uDSB [117]: Unbalanced Diffusion Schrödinger Bridge (UDSB) is an extension of the Diffusion Schrödinger Bridge (DSB) framework that allows for modeling the temporal evolution of populations with changing mass over time. Unlike traditional DSBs, which assume conservation of mass and work with probability measures, UDSBs can handle marginals with arbitrary finite mass. uDSBs achieve this by incorporating stochastic differential equations with killing and birth terms, and by deriving their time reversals. We use the Python codes on Github (`https://github.com/matteopariset/unbalanced_sb`) to run scNODE.

S2 Hyperparameter Tuning

To benchmark methods evaluated in this paper, we selected parameters based on the average 2-Wasserstein distance using 3-fold cross validation. Here, we split the cells in each training timepoint into 3 sets, and perform cross-validation such that 2 sets are used for training, and the third for testing. To search for optimal hyperparameters, we used *wandb* [13].

For baseline uDSB, we first used PCA to project the gene expression to 50-dimensional space. For the networks Q_θ and $\widehat{Q}_{\widehat{\theta}}$, the `x_encoder` was 3-layer MLP with 300-dimension hidden layers, the `t_encoder` was 3-layer MLP with 32 dimension hidden layers, with takes 16-dimensional sinusoidal embedding of time t, and, the `decoder` was a 3-layer MLP decoder with 300 dimension hidden layers accepts concatenation of outputs from `x_encoder` and `t_encoder`. The K_ω network includes a 5-layer MLP with 64-dimension hidden layers. As activation functions, the `SiLU` for Q_θ and $\widehat{Q}_{\widehat{\theta}}$, and, `Leaky-ReLU` for K_ω networks. The ADAM optimizer wth gradient clipping was used, and initial learning rates for Q_θ and $\widehat{Q}_{\widehat{\theta}}$ were set to 0.001, and 0.01 for K_ω. A batch size of 512 was used, and total training was conducted over 10 iterations, where each iteration included 10 epochs of forward and 10 epochs of backward training.

For PRESCIENT, we searched over the following hyperparameter spaces: latent dimension $\in \{10,50\}$, number of hidden layers $\in \{1,2,3\}$, sd $\in [0.0,1.0]$, tau $\in [0.0,0.1]$, gradient clipping $\in [0.0,1.0]$. The remaining hyperparameters were set

to default. We estimated the growth rates for the EMT [31] and pancreatic [152] datasets using the mean of z-scores annotated to birth (KEGG_CELL_CYCLE) and death (KEGG_APOPTOSIS) as suggested in [163].

For MIOFlow, we searched over the following hyperparameter spaces: gae embedded dim ∈ {10,50}, layers ∈ {[50,50],[16,32,16]}, λ ∈ [1,40,100]. The remaining hyperparameters were set to default.

For scNODE, we searched over the following hyperparameter spaces: latent dimension (d) ∈ {10,50}, encoder network size ∈ {None, [d], [d,d]}, decoder network size ∈ {None, [d], [d,d]}, drift network size ∈ {None, [d], [d,d]}.

For ARTEMIS, we searched over the following hyperparameter spaces: latent dimension ∈ {10,50}, vae encoder hidden dimension ∈ {[512,256], [256,128]}, vae decoder hidden dimension ∈ {[512,256], [256,128]}, vae pre-train epochs (E_1) ∈ {50,100}, number of iterations (E_2) ∈ {2,4,6,8,10}, all learning rates ∈ {0.001,0.0001}, vae batch size ∈ {32, 64}, SDE sampling batch size ∈ {256,512}.

S4 Investigate ARTEMIS hyperparameters

We next investigate the hyperparameters tuned in ARTEMIS and their effects on prediction performance at unmeasured timepoints. Using the three datasets [152, 46, 31], we evaluate its performance on the previously defined task on held out timepoints. We hope this analysis provides heuristic guidance to users to choose hyperparameters for training ARTEMIS. We vary the following hyperparameters:

1. base drift f from {0,2,4,8,10,50,100}: Figure A.19a. compares ARTEMIS's performance across different values of the base drift f. Lower values lead to better prediction performance, whereas higher values can dominate the learned forward and backward drifts, resulting in trajectories primarily driven by the base drift rather than the learned dynamics.

2. Number of discretization steps from {12,50,75,100}: As shown in Figure A.19b., the prediction performance remains largely consistent across different numbers of discretization steps. Therefore, users may select the number of steps based on the desired level of trajectory resolution.

3. VAE latent dimension d from [10,25,50,75,100,125,175,200]: Figure A.19c. compares the VAE latent dimensions. We find that a lower latent dimension (e.g., 10) is sufficient for simpler datasets. However, for more complex datasets with multiple cell types, we recommend using larger latent dimensions (>50) to better capture variability.

4. SDE sampling batch size from {32,64,128,256,512}" Figure A.19d. shows that a higher sampling size improves performance as it effectively penalizes more samples during training within a fixed number of epochs. However, a larger sampling size could cause computational overhead, thus, users can make a reasonable choice based on a tradeoff between accuracy and computational costs.

5. Effect of using L_{joint} loss terms, i.e., including either, both, or none of the terms: The results in Figure A.19e. indicate that including both losses enables joint optimization of the VAE and uDSB components, leading to better performance compared to using either loss individually or omitting them entirely.

S5 Identify drift-genes

To bridge the latent forward drift dynamics with gene expression changes, we map the learned latent drift values to the gene expression space. Let $\overrightarrow{z}_{t,i} \in \mathcal{R}^d$ be a latent variable generated using the forward SDE (Eq. 9a) for a cell i at time t. Let $\widehat{x}_{t,i} \in \mathcal{R}^g$ be the reconstructed gene expression for the cell i using the decoder (p_ϕ). To map the forward drift at t to the gene expression space, we multiply the output of $Q(\overrightarrow{z}_{t,i}, t; \theta) \in \mathcal{R}^d$ with the jacobian of p_ϕ to compute the jacobian vector product (JVP) [126].

The jacobian $J \in \mathcal{R}^{g \times d}$ of p_ϕ is given by:

$$J_i = \begin{bmatrix} \frac{\partial \widehat{x}_{t,i,(1)}}{\partial \overrightarrow{z}_{t,i,(1)}} & \frac{\partial \widehat{x}_{t,i,(1)}}{\partial \overrightarrow{z}_{t,i,(2)}} & .. & \frac{\partial \widehat{x}_{t,i,(1)}}{\partial \overrightarrow{z}_{t,i,(d)}} \\ . & . & . & . \\ . & . & . & . \\ \frac{\partial \widehat{x}_{t,i,(g)}}{\partial \overrightarrow{z}_{t,i,(1)}} & \frac{\partial \widehat{x}_{t,i,(g)}}{\partial \overrightarrow{z}_{t,i,(2)}} & .. & \frac{\partial \widehat{x}_{t,i,(g)}}{\partial \overrightarrow{z}_{t,i,(d)}} \end{bmatrix}, \tag{A.1}$$

Then, the drift scores for cell i can be calculated as:

$$\text{drift\_scores}_i = J_i Q(\overrightarrow{z}_{t,i}, t; \theta), \tag{A.2}$$

where $\text{gene\_drift\_scores}_i \in \mathcal{R}^g$. This is then averaged over all cells at time t to get an average score for each gene:

$$\text{drift\_scores} = \frac{\sum_{i=1}^{n_t} J_i Q(\overrightarrow{z}_{t,i}, t; \theta)}{n}, \tag{A.3}$$

where $\text{drift\_scores} \in \mathcal{R}^g$ and drift_scores_j gives a gene-drift-score for gene j. We used the `jax.jvp()` function from JAX library for computing forward-mode Jacobian-vector product.

Significance testing of drift-genes

To evaluate the significance of the gene-drift scores, we randomized the positions of cells in the latent space and calculated their forward drifts. The drift values from the randomized cells, combined with the decoded gene expression of the original unshuffled cells, were used to establish a null distribution for gene-drift scores. A two-sided t-test was then performed to compare the observed gene-drift scores against the null distribution. Most drift genes identified by ARTEMIS were found to be significant, with p-values < 0.05.

Drift gene bias toward highly expressed genes

To assess potential bias, we computed the Spearman correlation between drift scores and average gene expression across three datasets (Supplementary Figure S6a.). While some timepoints showed high correlation (> 0.5), several had lower correlations (< 0.5). We also compared the top 20 drift and highly expressed genes, finding low overlap (< 10 genes) at several timepoints and high overlap (> 10 genes) at some others (Supplementary Figure S6b.). In cases of high overlap, we found the genes to be biologically relevant, including progenitor genes at t=0,1 and SC-β/SC-EC branch-associated genes at t=7 in the pancreatic dataset [152], as well as key developmental regulators in zebrafish [46]. At timepoints with low overlap (< 10 genes), drift genes had significantly lower expression than highly expressed genes. These findings show that while some drift genes are highly expressed, many are not driven by high expression.

S6 Perturbation analysis

For the EMT dataset [31], *in silico* perturbations were introduced by modifying the scaled normalized expression of target genes to z-score values: less than 0 for underexpression (knockdowns) and greater than 0 for overexpression. The resulting perturbed gene expression profile was then input to the trained ARTEMIS model, where it is was projected to the latent space to generate a forward SDE up to time T. Perturbations were introduced with magnitudes of -25,-20,-15,-10,10,15,20,25.

We conducted 10 trials, where 2,000 cells sampled from the timepoint of perturbation introduction were used to predict cellular trajectory up to time T using the pre-trained ARTEMIS model. These simulations allowed us to examine the effects of perturbations within the latent space.

We trained a multilayer perceptron (MLP) classifier using the Python library scikit-learn [118] to predict timepoints based on latent cell representations. For a given trajectory resulting from perturbations introduced at time t, the predicted trajectory in the latent space was classified using the pre-trained MLP classifier.

Then, the number of predicted cells classified into each timepoint were compared between perturbed and unperturbed trajectories by performing a two-sided t-test test. We used this analysis to identify if the introduced perturbations for the drift-genes could alter/reverse the epithelial-to-mesenchymal transition by generating more cells corresponding to earlier or later timepoints (Figure 5, Supplementary Figure S4).

S7 Performance on sparse datasets

We further evaluated ARTEMIS on a mouse hematopoiesis dataset with lineage tracing information spanning three timepoints, comprising 49,302 cells [159]. To assess its predictive performance, we withheld the second time point. ARTEMIS demonstrated superior accuracy in reconstructing the held out time point (see Supplementary Table S4), indicating its robustness even in sparse settings. However,

a limited number of time points may reduce the resolution of inferred trajectories, particularly when key transitions occur between unobserved intervals.

Evaluating clonal fate prediction

We further evaluated ARTEMIS on lineage tracing data to assess its ability to predict clonal fate within neutrophil and monocyte lineages. Here, we trained ARTEMIS on all timepoints and initialized the trained model with cells from $t = 0$ and having clonal information across three timepoints to predict forward trajectories. Using the trained model, the inferred forward cell drift captured the differentiation lineages (see Supplementary Figure S9a.), recovered relative cell population changes, and predicted cell statuses (see Supplementary Figure S9b.).

For clonal bias analysis, cells at the final timepoint were classified as neutrophil, monocyte, or others using a nearest neighbor classifier trained on the groundtruth gene expression. Following the approach in [1], we computed clonal fate probability for each cell at $t = 0$as the fraction of its clonal relatives that became neutrophils or monocytes as ground truth. For ARTEMIS predicted trajectories, this was approximated as the number of neutrophils divided by the total number of monocytes and neutrophils within each cell's trajectory (see Supplementary Figure S9c.). ARTEMIS recovered clonal bias well for the neutrophil lineage and weakly for the monocyte lineage.

Data preprocessing

The mouse hematopoiesis dataset is available at `https://github.com/AllonKleinLab/paper-data`. We used cells from all three timepoints with lineage tracing information, log-normalized the training and held out time points separately, selected the top 2,000 highly variable genes, and removed cell cycle genes based on the training set, following the recommendations in [81]. We used the same strategy as in Supplementary Note S1 for evaluating prediction performance.

S8 Effect of cell population modeling in ARTEMIS performance

ARTEMIS models relative cell population changes using cell counts as ground truth rather than relying on prior knowledge, such as proliferation or apoptosis-associated genes. To assess the effect of cell population modeling, we trained a variant of ARTEMIS without the neural network for kill rate inference and compared its performance across three datasets for training and held out timepoints (Figure A.22). We also show a comparison with the PRESCIENT model trained with growth rates inferred from prior knowledge. Results show that ARTEMIS performs similarly with and without explicit cell population modeling, as the kill rate network is trained independently and can be excluded, allowing uDSBs to function as diffusion Schrödinger bridges (DSBs). However, incorporating cell population modeling enables ARTEMIS to infer cell population changes.

Supplementary Figures

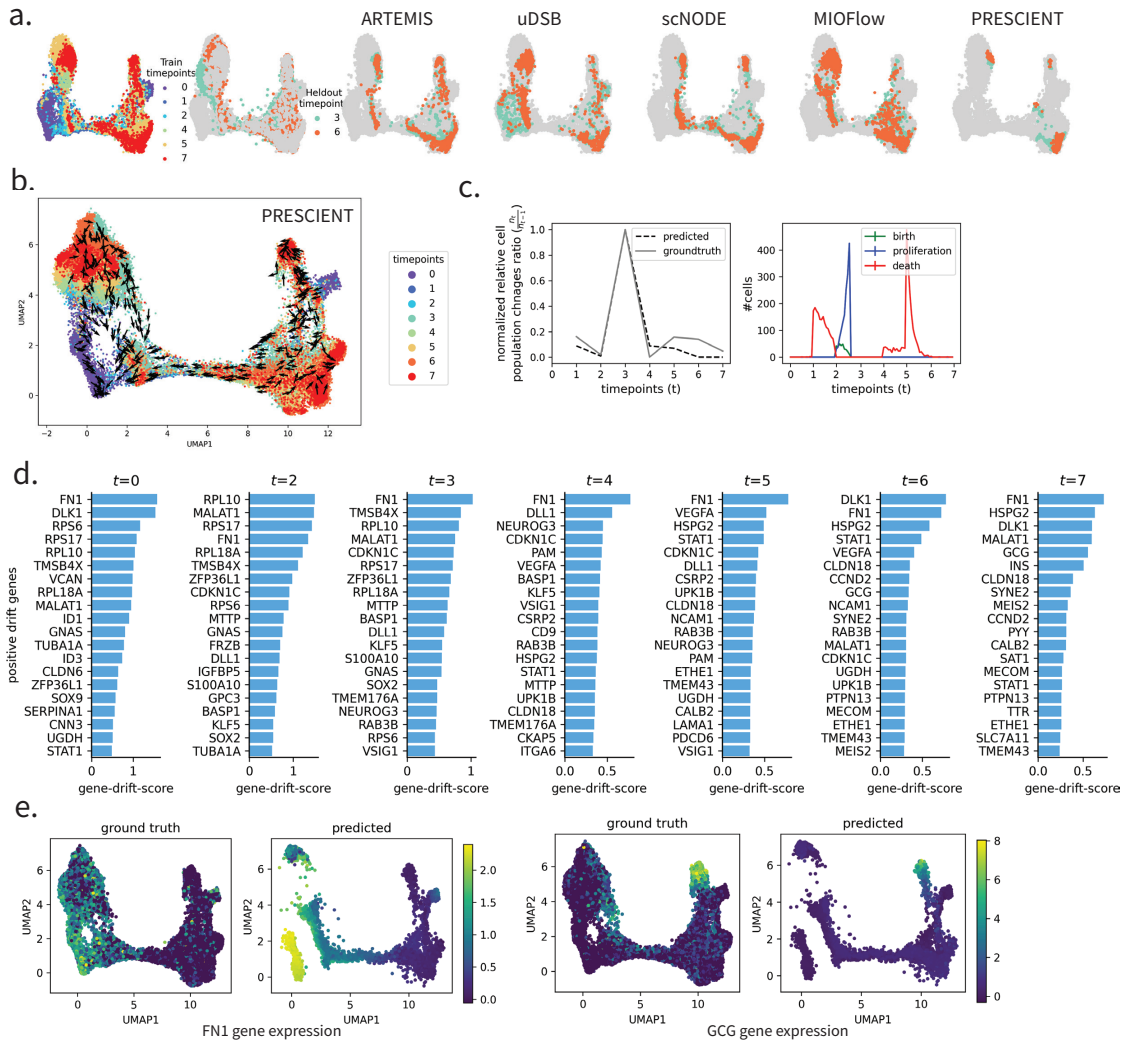

Figure A.15: **Application to β-cell differentiation in human pancreas.** a.) Benchmarking ARTEMIS against state-of-the-art methods for predicting gene expression at held-out timepoints (3,6) using pancreatic data. b.) Cell drift inferred by Prescient (w.o. growth rates). c.) Left: Comparison of normalized ratios of relative cell population changes between ground truth and ARTEMIS-predicted cell statuses as live, Right: Number of cells predicted as born, proliferated, and died throughout the trajectory. An increase in cell births and proliferation was observed at $t = 2$ and $t = 3$, coinciding with an increase in relative cell population in the ground truth data. Conversely, a significant number of cells were predicted to die between $t = 1$ to $t = 2$ and $t = 4$ to $t = 6$, aligning with the observed decline in relative cell population. d.)Drift genes identified for zebrafish dataset for remaining ten timepoints. e.) Ground truth vs. predicted gene expression of drift genes FN1 (identified across all timepoints) and GCG (identified at t=6,7).

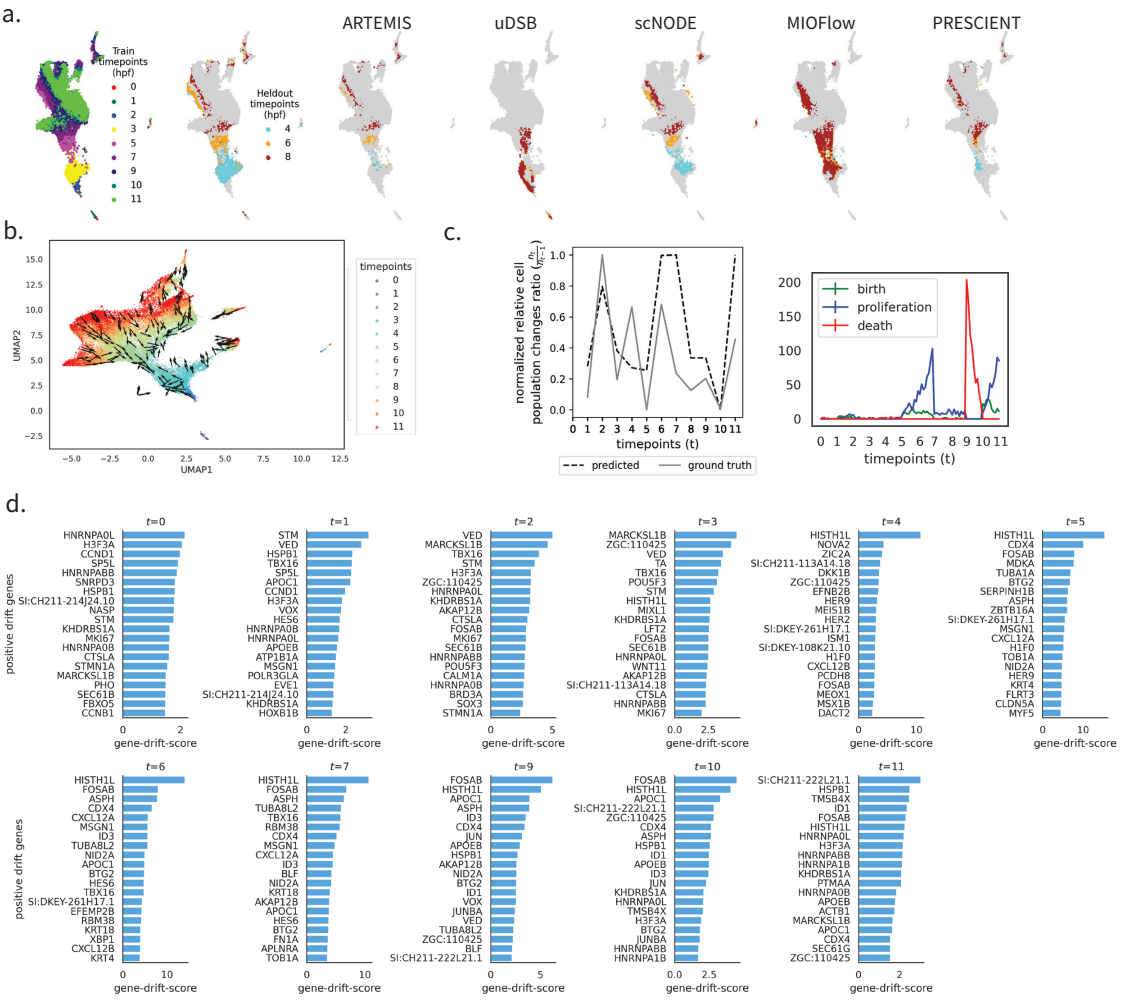

Figure A.16: **Application to Zebrafish embryogeneis data.** a.) Benchmarking ARTEMIS against state-of-the-art methods for predicting gene expression at held-out timepoints (4, 6, 8) using zebrafish data. b.) Cell drift inferred by PRESCIENT. c.) Left: Comparison of normalized ratios of relative cell population changes between ground truth and ARTEMIS-predicted cell statuses as live, Right: Number of cells predicted as born, proliferated, and died throughout the trajectory. An increase in cell births and proliferation was observed between $t = 4$ to $t = 7$ and $t = 10$ to $t = 11$, coinciding with an increase in relative cell population in the ground truth data. Conversely, a significant number of cells were predicted to die between $t = 9$ to $t = 10$, aligning with the observed decline in relative cell population. d.) Drift genes identified for zebrafish dataset for remaining ten timepoints.

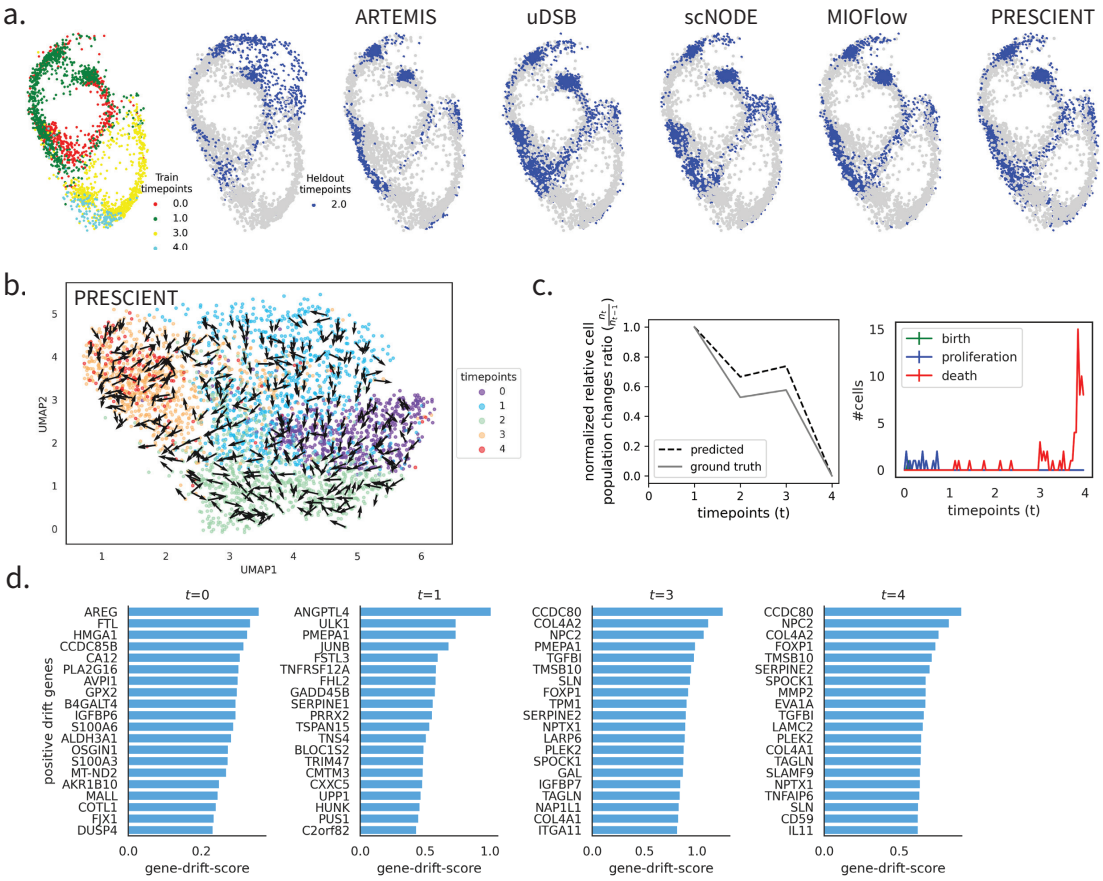

Figure A.17: **Application to A549 lung cancer cells treated with *TFGB1* to induce EMT.** a.) Benchmarking ARTEMIS against state-of-the-art methods for predicting gene expression at held-out timepoint (2) using EMT data. b.) Cell drift inferred by PRESCIENT. c) Left: Comparison of normalized ratios of relative cell population changes between ground truth and ARTEMIS-predicted cell statuses as live, Right: Number of cells predicted as born, proliferated, and died throughout the trajectory. Shorter intervals of predicted cell proliferation are observed earlier in the trajectory, potentially reflecting the higher cell numbers in the ground truth data at these early timepoints. Cell death events are distributed across several shorter intervals throughout the trajectory. A significant spike in predicted cell death is observed towards the end of the trajectory, coinciding with a substantial reduction in the relative cell population in the ground truth data. d.) Drift genes identified for emt dataset for remaining four timepoints.

Figure A.18: **Perturbation results for different levels of over and underexpression of genes** (a) *TPM1* and (b) *AMIGO2*: (-20,-15,-15,-10,0,10,15,20) and additionally (-25,25) for *AMIGO2*. Cells are assigned to specific timepoints by an MLP classifier, and the number of cells generated from perturbed and unperturbed trajectories were compared using a two-sided t-

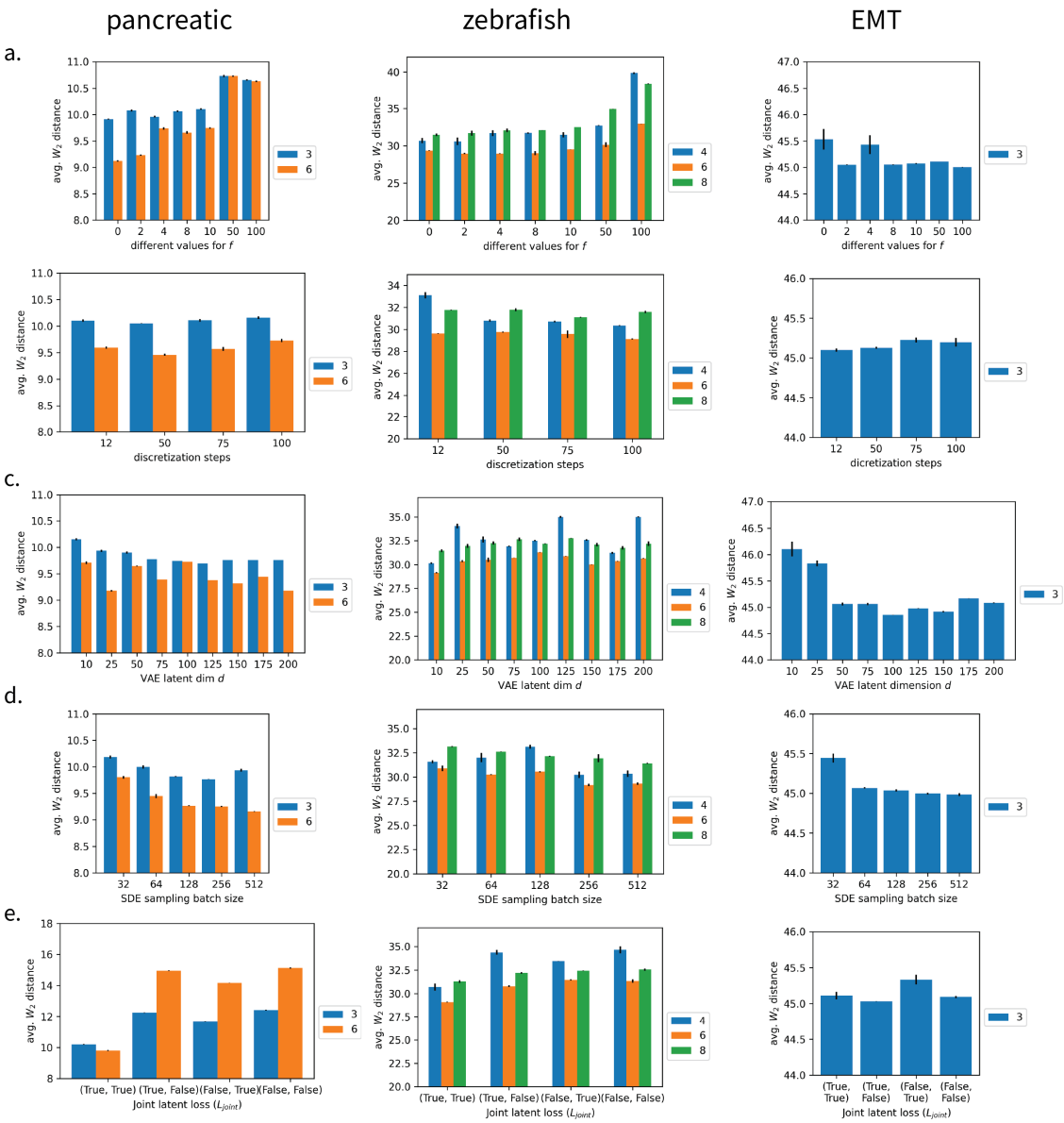

Figure A.19: **Evaluating different hyperparameter sets and training configurations for ARTEMIS.** The different hyperparameters/configurations are: a.) base drift f, b.) discretization steps, c.) VAE latent dimension$d$, d.) SDE sampling batch size, e.) effect of using $L_{joint} = W_2(Z_{\varphi,t}, \overrightarrow{Z}_t) + W_2(X_t, p_\phi(Z_{\varphi,t}))$ loss terms. The x-axis $(\cdot,\cdot)$ indicates when either of the W_2 losses is used.

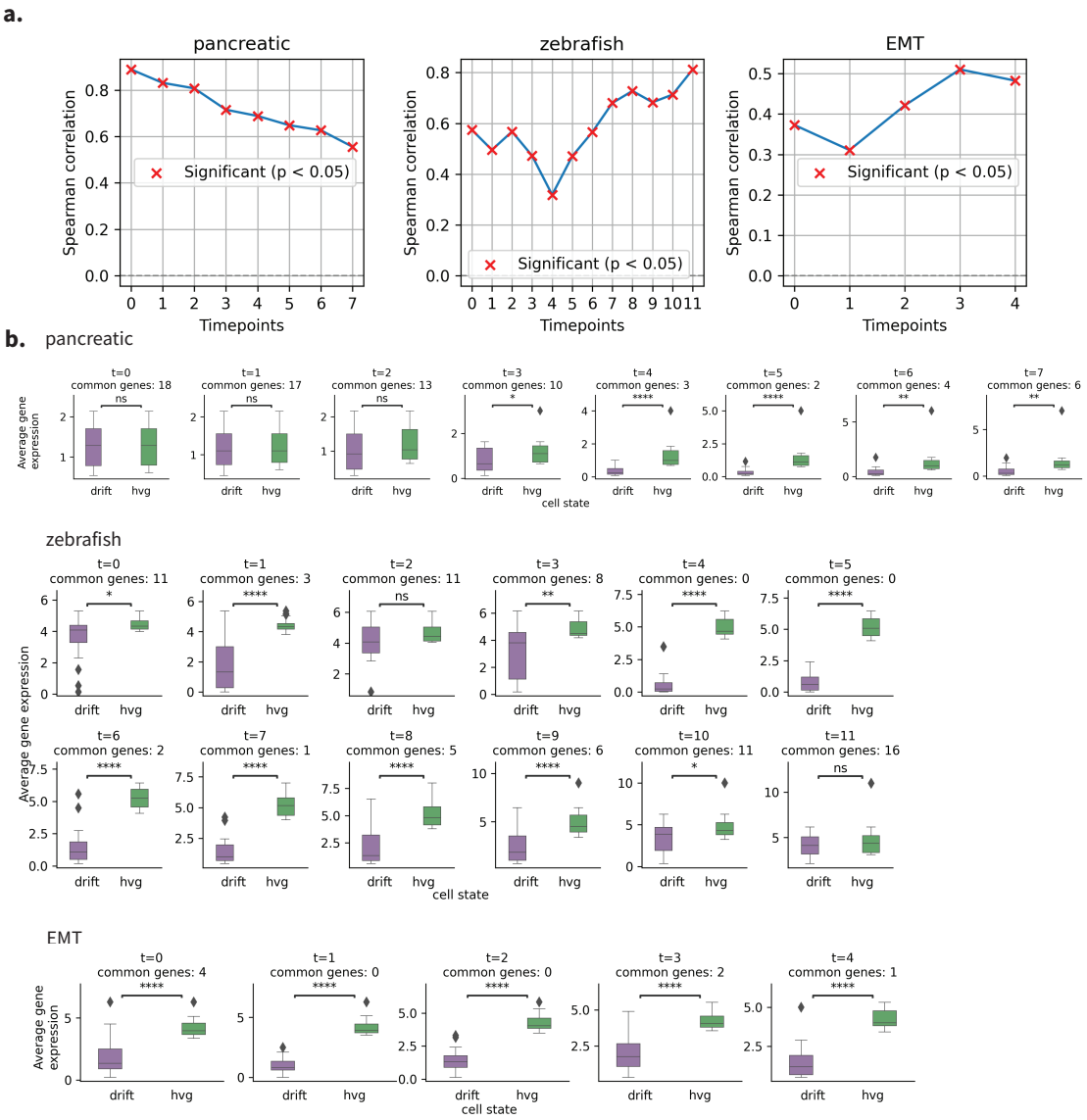

Figure A.20: a.) Spearman correlation between drift gene scores and average gene expression to assess potential bias toward highly expressed genes. b.) Comparison of expression levels between the top 20 drift genes and the top 20 highly expressed genes at each timepoint, including the number of common or overlapping genes between the two sets.

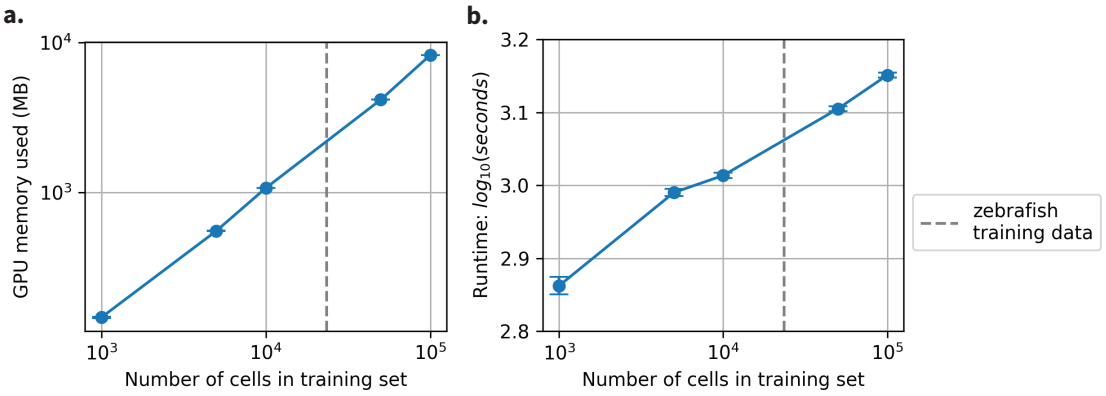

Figure A.21: a.) memory and b.) runtime estimates for increasing training set sizes tested on the NVIDIA RTX A6000 GPU using the zebrafish dataset. The hyperparameters used in training are: base drift f=2, VAE latent dim=10, SDE sampling batch size=256, discretization steps=100, VAE pre-train epochs=50.

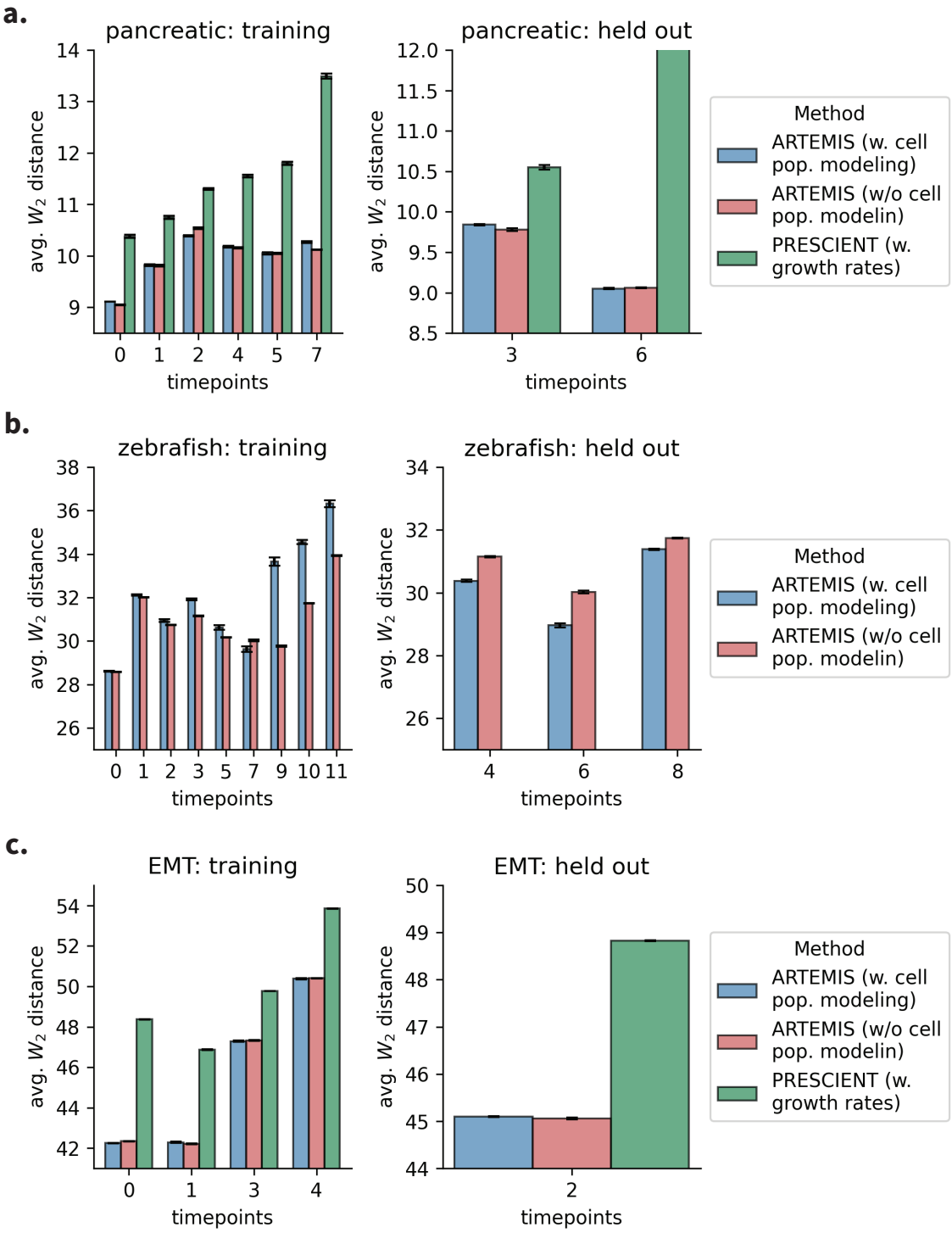

Figure A.22: a.- c.) Training and held out performance comparing ARTEMIS trained with and without cell population modeling across three datasets, also compared to Prescient trained with growth rates from prior knowledge.

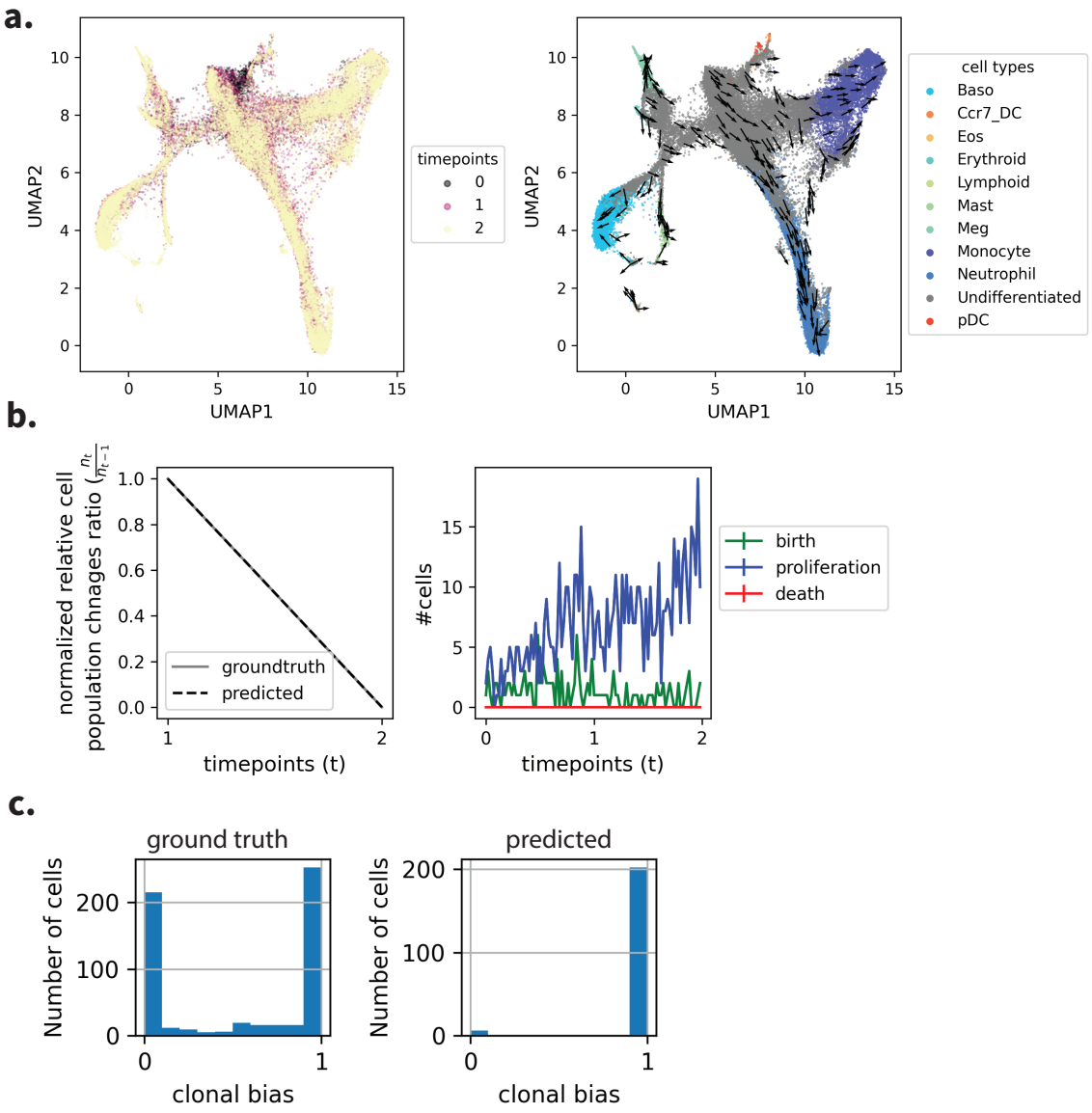

Figure A.23: **Application to mouse hematopoiesis.** a.) Left: ground truth data colored by time. Right: Visualization of the drift inferred by ARTEMIS trained on all timepoints. b.) Left: Comparison of normalized ratios of relative cell population changes between ground truth and ARTEMIS-predicted cell statuses as live. Right: Number of cells predicted as born, proliferated, and died across the trajectory. Cell proliferation increases throughout the differentiation landscape as the cell population increases in the ground truth data. The rate of cell proliferation and birth is higher from $t = 0$ to $t = 1$ compared to that between $t = 1$ to $t = 2$, as the increase in cell population from $t = 1$ to $t = 2$ is 2 times compared to that from $t = 0$ to $t = 1$, which was 3.2 times. c.) The distribution of ground truth and predicted clonal fate bias for neutrophil and monocyte lineages is computed as the number of neutrophils divided by the total number of monocytes and neutrophils within each cell's trajectory.

Supplementary Tables

Table A.30: Wasserstein distance between the predicted and training timepoints for the pancreatic dataset [152]. Numbers in **bold** indicate best performance.

Method	t=0	t=1	t=2	t=4	t=5	t=7
ARTEMIS	**9.11±0.00**	**9.82±0.02**	**10.39±0.01**	**10.18±0.02**	**10.05±0.02**	**10.27±0.02**
uDSB	9.38±0.002	10.81±0.01	11.89±0.01	11.94±0.01	11.94±0.01	11.94±0.001
scNODE	9.5±0.001	10.08±0.001	10.54±0.00	10.35±0.001	10.21±0.001	10.29±0.001
MIOFlow	12.22±0.01	11.73±0.01	11.73±0.01	11.49±0.01	11.58±0.02	12.12±0.02
Prescient	10.38±0.03	10.73±0.02	11.22±0.01	11.29±0.01	11.38±0.01	12.48±0.02
Prescient (w. growth rates)	10.38±0.03	10.75±0.03	11.30±0.02	11.55±0.03	11.80±0.03	13.49±0.05

Table A.31: Wasserstein distance between the predicted and training timepoints for the zebrafish dataset [46]. Numbers in **bold** indicate best performance.

	t=0	t=1	t=2	t=3	t=5
ARTEMIS	**28.6±0.021**	**32.11±0.03**	**30.94±0.07**	**31.92±0.04**	**30.62±0.11**
uDSB	29.27±0.009	34.08±0.05	37.18±0.2	42.5±0.22	44.87±0.18
scNODE	29.66±0.002	32.72±0.001	31.75±0.002	32.17±0.002	30.76±0.003
MIOFlow	46.42±0.02	43.75±0.02	35.33±0.02	34.68±0.01	33.90±0.02
Prescient	61.35±0.2	61.71±0.22	54.47±0.21	53.04±0.2	51.80±0.2

	t=7	t=9	t=10	t=11
ARTEMIS	**29.62±0.13**	33.66±0.19	**34.55±0.09**	36.32±0.16
uDSB	42.88±0.07	44.96±0.06	45.82±0.08	46.91±0.08
scNODE	30.55±0.003	**33.66±0.002**	35.12±0.001	**36.06±0.002**
MIOFlow	33.11±0.04	37.22±0.04	39.55±0.04	41.79±0.05
Prescient	48.73±0.26	52.99±0.26	56.61±0.25	58.71±0.26

Table A.32: Wasserstein distance between the predicted and training timepoints for the EMT dataset [31]. Numbers in **bold** indicate best performance.

	t=0	t=1	t=3	t=4
ARTEMIS	**42.25±0.01**	**42.29±0.03**	**47.29±0.03**	**50.38±0.03**
uDSB	42.28±0.05	44.18±0.12	49.64±0.12	52.34±0.10
scNODE	42.75±0.002	42.70±0.003	49.7±0.01	52.45±0.02
MIOFlow	44.98±0.02	42.96±0.01	49.68±0.02	56.08±0.06
Prescient	45.31±0.03	44.33±0.05	47.99±0.03	51.60±0.01
Prescient (w. growth rates)	48.37±0.02	46.88±0.02	49.77±0.01	53.85±0.01

Table A.33: Wasserstein distance between the predicted and held-out timepoints for mouse hematopoiesis dataset. Numbers in **bold** indicate best performance.

Timepoints	**train**		**held out**
	t=0	t=2	t=1
ARTEMIS	3.69±0.015	4.99±0.019	4.62±0.04
uDSB	3.86±0.02	5.7±0.04	6.25±0.05
scNODE	3.82±0.0001	**4.94±0.0002**	**4.54±0.0005**
MIOFlow	**3.66±0.0002**	5.46±0.011	4.87±0.009
Prescient	4.8±0.003	7.2±0.004	5.8±0.003

A.3 iBrainMap: Supplement

Supplementary Notes

Supplementary Note 1: Graph Diffusion

We applied network diffusion to propagate well-known disease genes throughout the Personalized Functional Genomics graphs (PFGs) through their edges (Figure A.24). Here we use the concept of insulated heat diffusion[93], where disease genes are treated as heat sources that retain some heat β ($0 < \beta < 1$) and equally distribute the rest to their neighboring nodes. The amount of heat propagated by disease genes, called diffused scores, represents the impact of these genes and is used as edge weights for the PFGs. The network diffusion requires two inputs: PFGs and prior knowledge (disease gene lists).

Let $G_i = (V_i, E_i)$ be the PFG for donor i, with $N_i = |V_i|$ nodes and E_i representing the edges. Let $A_i \in R^{N_i \times N_i}$ be the adjacency matrix and $D_i \in R^{N_i \times N_i}$ be its diagonal matrix of the out-degree of nodes. The diffusion matrix F_i is calculated as:

$$F_i \;=\; \beta\Big(I - (1-\beta)\, A_i\, D_i^{-1}\Big)^{-1}. \tag{A.4}$$

Given a collection of disease genes d_g, let $P_i \in R^{N_i \times N_i}$ be the diagonal encoding prior knowledge for a donor i:

$$(P_i)[j,j] = \begin{cases} 1, & \text{if gene j is in } d_g, \\ 0, & \text{otherwise,} \end{cases} \tag{A.5}$$

We combine the PFG and the prior knowledge to define the final edge weights for G_i as

$$M_i = F_i\, P_i. \tag{A.6}$$

We compute two versions of M_i : $M_{AD,i}$, $M_{SCZ,i}$ using our identified gene lists for AD and SCZ, respectively. We refer to M_{AD} and M_{SCZ} as bio-diffused PFGs, which are used to train our graph neural network model. We can also compute these matrices for other gene-of-interest (GOI) lists (M_{GOI}). In this paper, we focus on AD and SCZ gene lists. Additionally, we define a third PFG, $M_{data-driven,i} = A_i \in R^{N_i \times N_i}$, which is the adjacency matrix of G_i and imbues no additional prior information.

Supplementary Note 2: Knowledge-guided Graph Neural Network (KG-GNN) Architecture

KG-GNN uses a self-attention mechanism to calculate attention scores between a node and its neighbors (Figure A.25). These are normalized across the node's neighborhood using a softmax function and further used to update the node's features. This helps assign different weights to different neighbors, offering higher flexibility and interpretability to the model. For a Personalized Functional Genomics graph (PFG) G_i (Supplementary Note 1, Figure A.24), we define a set of node features

$h^l = \{h_1^l, h_2^l, ..., h_{N_i}^l\}$, where $h_t^l \in R^{d_i}$ represents the feature vector for node t at layer l, and d_i is the constant feature dimension. The node features h^l represent the main model output and are initialized as average gene expression of 2,766 highly variable genes (HVGs) for cell type nodes and the coexpression of the HVGs with gene expression for the TF/TG nodes, as documented in the definition of PFGs section of the main manuscript, then refined through model iterations. All node features and KG-GNN computations are donor-specific (i), but for simplicity, we omit i in notation when discussing an individual donor where it is clear from context. However, keep in mind that each donor has a distinct graph and feature set.

The unnormalized attention coefficient $e_{t,s}^l$ between any node t and its neighboring node s is:

$$e_{t,s}^l = \text{LeakyReLU}\Big(a^l\big(W^l h_t^l \,\|\, W^l h_s^l\big)\Big), \tag{A.7}$$

, where $W^l \in R^{d_{l+1} \times d_i}$ is a learnable weight matrix which acts as an 'encoder' to compare features between nodes, $\alpha^l \in R^{2d_{l+1}}$ is a learnable weight vector, and $||$ denotes vector concatenation. We then compute the normalized attention coefficient $\alpha_{t,s}^l$ by normalizing $e_{t,s}^l$ across the node t's neighborhood Ω_t using a softmax function:

$$\alpha_{t,s}^l = softmax(e_{t,s}^l) = \frac{\exp(e_{t,s}^l)}{\sum_{n \in \Omega_t} \exp(e_{t,n}^l)}. \tag{A.8}$$

, where $\sum_{s \in \Omega_t} \alpha_{t,s}^l = 1$.

Before we update the node feature h_t^l , we incorporate prior biological knowledge through edges using our bio-diffused PFGs M_{AD}, M_{SCZ} to get bio-diffused attention scores:

$$b_{D,t,s}^l = \frac{M_D[t,s]\alpha_{t,s}^l}{\sum_{n \in \Omega_t} M_D[t,n]\alpha_{t,n}^l}. \tag{A.9}$$

where D denotes AD, SCZ, or data-driven priors. Specifically, the bio-diffused attention scores for each prior are:

$$\begin{aligned} b_{AD,t,s}^l &= \frac{M_{AD}[t,s]\alpha_{t,s}^l}{\sum_{n \in \Omega_t} M_{AD}[t,n]\alpha_{t,n}^l}, \\ b_{SCZ,t,s}^l &= \frac{M_{SCZ}[t,s]\alpha_{t,s}^l}{\sum_{n \in \Omega_t} M_{AD}[t,n]\alpha_{t,n}^l}, \\ b_{data-driven,t,s}^l &= \frac{M_{data-driven}[t,s]\alpha_{t,s}^l}{\sum_{n \in \Omega_t} M_{data-driven}[t,n]\alpha_{t,n}^l}, \end{aligned} \tag{A.10}$$

, where $\sum_{s \in \Omega_t} b_{D,t,s}^l = 1$.

Finally, the updated node feature for the node t is computed by aggregating attention scores and passing it through a non-linearity σ (sigmoid) over its neighbors:

$$h_t^{l+1} = \sigma\Big(\sum_{s \in \Omega_t} b_{D,t,s}^l \, W^l \, h_s^l\Big), \tag{A.11}$$

For multi-head attention, the above operations are replicated K times (each with different parameters), and the output is aggregated by adding feature-wise:

$$h_t^{l+1} = \sigma\Big(\frac{1}{K}\sum_{k=1}^{K}\sum_{s\in\Omega_t} b^{l,k}_{D_k,t,s}\, W^{l,k} h_s^l\Big), \tag{A.12}$$

where D_k is the prior associated with head k. For our particular use case, we divided the number of heads K into AD-driven (K_{AD}), SCZ-driven (K_{SCZ}), and data-driven ($K_{data-driven}$), $K = K_{AD} + K_{SCZ} + K_{data-driven}$, and updated h_t^{l+1} as follows:

$$\begin{aligned} h_t^{l+1} = \sigma(\frac{1}{K}(&\sum_{k=1}^{K}{}_{AD}\sum_{s\in\Omega_t} b^{l,k}_{AD,t,s} W^{l,k} h_s^l \\ &+ \sum_{k=1}^{K}{}_{SCZ}\sum_{s\in\Omega_t} b^{l,k}_{SCZ,t,s} W^{l,k} h_s^l \\ &+ \sum_{k=1}^{K}{}_{data-driven}\sum_{s\in\Omega_t} b^{l,k}_{data-driven,t,s} W^{l,k} h_s^l)) \end{aligned} \tag{A.13}$$

This process is repeated with unique h_l, M_D and shared $W^{l,k}$, α^l for all donors in each epoch. Once trained, the model outputs latent graph embeddings $z_i \in R^{d^L}$, where d^L is the dimension of the final L^{th} layer, for each bio-diffused PFG G_i for a donor i. This is computed by averaging all the node features in the L^{th} layer of KG-GNN:

$$z_i = \frac{1}{N_i}\sum_{t=1}^{N_i} h_t^L, \tag{A.14}$$

The KG-GNN model optimizes the following binary cross-entropy loss across all j donors:

$$\mathcal{L} = \frac{1}{J}\sum_{i=1}^{J}\Big(y_i \log(\mathrm{MLP}(z_i)) + (1-y_i)\log\big(1-\mathrm{MLP}(z_i)\big)\Big), \tag{A.15}$$

where we train a Multi-Layer Perceptron (MLP) to classify z_i into AD vs. control class.

Supplementary Note 3: Glossary of Terms

Our analysis includes multiple phenotype contrasts split into three levels: Disease vs. control, Disease Progression, and Neuropsychiatric Symptoms. This section provides their definitions.

Disease vs. control

1. AD vs. control This contrast compares donors with AD and controls. Donors with AD are defined as those who have CERAD scores of 2, 3, or 4, BRAAK stage of 3 and above, and clinically proven dementia. The control group is donors with a CERAD score of 1 and BRAAK stage within 0-3.

2. SCZ vs. control SCZ is any donor with SCZ diagnosis (SCZ — Schizoaffective_bipolar — Schizoaffective_depressive) and without a secondary diagnosis, except for metabolic and eating disorders.

3. AD-DLBD vs. control DLBD is any donor with a DLBD diagnosis (DLBD), and a secondary diagnosis can be only AD.

4. Pathology-cognition (AD-resilient vs. AD-strict vs. control) This contrast integrates pathological and cognitive information to group donors based on CERAD score, BRAAK score, and Cognitive Resilience score. The "Cognitive Resilience" score is just a simple residual with Clinical Dementia Rating (CDR) sum of boxes as the outcome regressed on Age, Sex, PMI, CERAD, and BRAAK_AD. CDR sum of boxes (CDR_SumBoxes) is the Clinical Dementia Rating Sum of Boxes and is defined as the sum of CDR domain values (memory, orientation, judgement, community, home and hobbies, and personal care)[115]. AD resilience has three groups:

   - AD-strict includes donors with CERAD = Definite AD, BRAAK stages ¿ 3, and clinically proven dementia.
   - Control group includes donors with CERAD = normal and BRAAK stage within 0-3
   - AD-resilient includes donors with the following conditions: a. CERAD = possible AD and CDR_SumBoxes MUST BE 0. b. CERAD = probable AD and CDR_SumBoxes MUST BE ¡=2 c. CERAD = definite AD and CDR_SumBoxes MUST BE ¡=3.

   For further details, please refer to `https://www.synapse.org/Synapse:syn26720956`.

Disease Progression

1. BRAAK This contrast compares AD progression via BRAAK stages that measure neurofibrillary tangles, irrespective of donors' clinical diagnosis.

2. Cognitive Dementia Rating Score This contrast compares the clinical dementia rating score (CDR score) where (0, 0.5, 1) = control, (2, 3) = Mild Cognitive Impairment (MCI), and (4, 5) = Dementia. For SEA-AD[52], we renamed their Cognitive status phenotype as follows: No Dementia = control, Dementia = Dementia.

3. CERAD This contrast compares AD progression via qualitative variables from neuropathological scoring, where 1 = no AD, 2 = possible AD, 3 = probable AD, and 4 = definite AD. For SEA-AD, we renamed their CERAD phenotypes

as follows: absent = no AD, sparse = possible AD, moderate = probable AD, Frequent = definite AD.

Neuropsychiatric Symptoms

1. Depression/Dysphoria vs. Control This contrast corresponds to depression and mood dysphoria. A donor is considered to be 'Case' if only mood dysphoria appears to be true and 'control' if mood dysphoria is not true and all other NPS corresponding to depression and mood are either NA or not true.

2. DecInt vs. control This contrast corresponds to depression and anhedonia. A donor is considered to be 'Case' if only anhedonia appears to be true and 'control' if anhedonia is not true and all other NPS corresponding to depression and mood are either NA or not true.

3. Sleep/WeightGain/Guilt/Suicide vs. control This contrast corresponds to sleep issues (early-, mid-, and late-insomnia, and hypersomnia), weight gain, guilt, and suicidal thoughts within AD lenient donors. A donor is considered to be 'Case' if at least one of the above symptoms appears to be true and 'control' if none of the symptoms are true.

4. Depression/Mood vs. control This contrast corresponds to depression and mood disorders. A donor is considered to be 'Case' if at least one of the above symptoms appears to be true and 'control' if none of the symptoms are true.

Supplementary Note 4: Training, Testing, and Validation

Graph Subsampling: We use the graph sampling technique, Neighbor Sampling[64], to sub-sample PFGs for training the KG-GNN model. In particular, we used the $Neighborloader$ function from PyTorch Geometric[47] and set the parameter $num\_neighbors$ to 10 neighbors to be sampled for each node for 100 iterations. This ensures connectivity of the subgraphs and information flow throughout the network. We set the $batch\_size$ (used for mini-batching) based on a hyperparameter to specify the number of subgraphs:

$$batch = \frac{numberofPFGs}{numberofsubgraphs} \tag{A.16}$$

We trained and tested our KG-GNN model on donors from the MSSM cohort for binary classification of AD vs. control. Here, we stratified split donors into 80% training and 20% held-out sets. We performed a 5-fold cross-validation (CV) to find the optimal hyperparameters and evaluated our model's performance on the held-out set. We also tested our pre-trained model on an independent dataset from the RADC cohort.

Cross-Validation: We adopted a modified cross-validation (CV) and testing scheme based on the MD-AD model5. The dataset was first divided into five equal parts. In each of five rounds, one part was held out as a test set, while the remaining

four parts were used for model tuning and training. Within each training set (4/5 of the data), we performed a five-fold CV to select the best hyperparameters based on prediction performance. The final model, trained on the entire training set using these selected hyperparameters, was then evaluated on the held-out test set. This process was repeated so that each part of the data served once as the test set.

Hyperparameter Tuning: We tuned our model over a range of hyperparameters: optimizer [Adam, SGD, Adagrad], learning rate $\in [1e-4, 5e-4, 1e-3, 5e-3]$, weight decay $\in [0, 5e-3, 5e-4]$, dropout $[0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8]$, batch size $[5, 10, 15]$, diffusion parameter $\in [0.1, 0.3, 0.5, 0.7, 0.9]$, number of subgraphs $\in [2, 3, 5, 10]$ and number of attention heads $\in [2, 4, 6, 8]$, number of GAT layers in KG-GNN model $\in [1, 2, 3]$, GAT input layer dimension $\in [2048, 1024]$, hidden GAT layer dimensions $\in [2048, 1024, 512, 256, 128, 64]$, MLP hidden dimensions $\in [128, 64]$. We show our model's performance under various hyperparameters in Figure A.27.

Performance metrics: We evaluated our model's performance using metrics such as balanced accuracy (BACC) and the area under the receiver operating characteristic curve (AUROC) to address the imbalanced nature of our training data. The following terms are used to compute BACC and AUROC:

$$BACC = \frac{sensitivity + specificity}{2} \tag{A.17}$$

$$sensitivity = TPR = \frac{TP}{TP + FN} \tag{A.18}$$

$$specificity = \frac{TN}{FP + TN} \tag{A.19}$$

$$FPR = \frac{FP}{FP + TN}, \tag{A.20}$$

where TP is true positive, TN is true negative, FP is false positive, FN is false negative, TPR is true positive rate, and FPR is false positive rate.

Our final model was picked based on the best average performance on the held-out test set based on the BACC and AUROC metrics for the AD classification task. Table A.34 reports the final model layers, hyperparameters, and training details.

Supplementary Note 5: Benchmarking machine learning and state-of-the-art algorithms for AD versus Control classification

We compared our knowledge-guided graph neural network (KG-GNN) model with other state-of-the-art graph learning algorithms like graph attention network (GAT) and graph convolutional networks (GCN) with the personalized functional genomics graphs (PFGs) as inputs for the AD vs. control classification task (Figure A.27c). GAT and GCN achieved BACC = 0.70 and AUC = 0.85, and BACC = 0.84 and AUC = 0.88, respectively. Additionally, we also benchmarked other machine learning algorithms like Support Vector Machine (SVM), Logistic Regression (LR), and

Multi-Layer Perceptron (MLP) to classify average cell type gene expression for each donor into AD vs. control groups (Figure A.28a). We ran these models using default settings from the Python package scikit-learn[45]. We also benchmarked KG-GNN with two state-of-the-art graph embedding methods: scGNN[157] and SIMBA[25]. We extracted cell-level embeddings for both methods and averaged them across donors to generate donor-level embeddings for AD vs. control classification. On held-out samples, scGNN achieved BACC = 0.57 and an AUC = 0.63, while SIMBA achieved BACC = 0.650 and AUC = 0.645. In comparison, iBrainMap graph embeddings achieved higher performance (BACC = 0.87, AUC = 0.91). Overall, the performance of these state-of-the-art models on held-out and independent datasets was lower than iBrainMap (Figure A.28b

Supplementary Note 6: KG-GNN Ablation Studies

To identify key contributors to KG-GNN's performance, we varied i) the number of genes used in constructing personalized functional genomics graphs, ii) diffusion strategies, and iii) biological networks, and tested performance on held-out and independent datasets. We found that increasing the number of genes improves performance up to a certain threshold (up to 10% TF, 10% TG), after which it declines (Figure A.29a). For diffusion strategies, random diffusion resulted in the lowest performance (Held-out: BACC = 0.813, AUC = 0.864; Independent: BACC = 0.588, AUC = 0.572), no diffusion performed better (Held-out: BACC = 0.85, AUC = 0.925; Independent: BACC = 0.537, AUC = 0.623), and our diffusion strategy achieved the highest performance (Held Out: BACC =0.87, AUC = 0.91; Independent: BACC =0.66, AUC = 0.808; Figure A.29b). We also evaluated alternate biological networks by constructing donor-specific co-expression networks using Dozer[104] (Figure A.29c). We used two types of node features: node2vec[57] and cell type gene expression, alongside two graph learning architectures: graph convolutional networks (GCN) and graph attention networks (GAT) for AD vs. control classification. Among these, the co-expression network model with cell type gene expression features and GAT architecture achieved the best performance (Held-out: BACC = 0.75, AUC=0.84; Independent: BACC = 0.363, AUC=0.33), which was lower than our KG-GNN performance (Figure A.29c).

Supplementary Note 7: Classifying graph embeddings across AD phenotypes

Classification for AD, SCZ, and related NPS We classified the graph embeddings into different AD phenotypes using machine learning algorithms like Support Vector Machines (SVM), Logistic Regression (LR), and Multi-layer Perceptron (MLP) using the Python package Scikit-learn[45]. To do this, we performed 5-fold cross-validation where the dataset was stratified split into training and held-out for each fold using a 4:1 ratio. We evaluated the performance using metrics like AUROC and BACC and picked the best model based on average AUROC across five folds. The results are shown in Figure 4.3b across several phenotypes for binary and multi-class classification tasks for donors from the MSSM cohort (Figure 4.2c): (1)

Binary classification: includes phenotypes SCZ, Dysphoria, DecInt, S/WG/G/S, D/M; (2) Multi-class classification: includes phenotypes AD-Resilience, BRAAK (early vs. mid vs. late stages), CERAD (No AD vs. Possible AD vs. Probable AD vs. Definite AD), Cogdx (Dementia vs. MCI vs. controls).

Classification for AD-Resilience & AD-DLBD Classifying graph embeddings from the AD resilience contrast is a three-class classification task with imbalanced samples across control, AD Resilience, and AD Strict. To assess the impact of class imbalance on performance, we reported key performance metrics, including AUC, balanced accuracy (BACC), and F1-score (Figure A.30a). We also attempted to classify AD-DLBD into two classes, AD and AD with DLBD. However, the model's performance was not very high (Figure A.30b), likely due to high class imbalance and the inherent complexity of comorbid conditions like DLBD with AD.

Supplementary Note 8: Node importance score computation

We use edge importance scores to derive node importance scores. For a node v_i, let N_{in} be the indegree of the node N_{out} and be the outdegree. We first calculate the indegree and outdegree importance scores of the node using the following formula:

$$I_{\text{in}} = \ln\Big(1 + \sum_j N_{\text{in}}\, b_{k,j,i}\Big), \tag{A.21}$$

$$I_{\text{out}} = \ln\Big(1 + \sum_j N_{\text{out}}\, b_{k,i,j}\Big), \tag{A.22}$$

$$\tag{A.23}$$

where $b_{k,j,i}$ and kk, i, j are incoming and outgoing importance scores of an edge between nodes v_i and v_j, respectively, and k can be AD-driven, SCZ-driven, or data-driven attention heads. Then the importance score of a node is computed using the formula:

$$\text{importance\_score}_t = \lambda\, I_t^{\text{in}} + (1-\lambda)\, I_t^{\text{out}}, \tag{A.24}$$

where $\lambda \in [0, 1]$ is a parameter to balance the indegree and outdegree importance scores. We empirically set $\lambda = 0.3$ based on the weighted average of the average indegree and outdegree of all nodes to give equal importance to both incoming and outgoing importance scores.

Supplementary Note 9: Cross-modal imputation and classification of graph embeddings

We developed a pipeline that imputes graph embeddings from genotype data and then uses the imputed embeddings for downstream classification tasks. Specifically, we trained an imputation model using genotype data to impute graph embeddings generated by a pre-trained KG-GNN model in the MSSM cohort. For the

imputation step, we evaluated four methods: CMOT[4], JAMIE[29], MOFA+[9], and autoencoder[10]. Once the imputed graph embeddings were obtained, we used them to classify disease labels using three different classifiers: random forest (RF), multi-layer perceptron (MLP), and support vector machines (SVM). We applied this pipeline to two classification tasks: ROSMAP [38] – classifying early (n = 88) versus late BRAAK stages (n = 113). ADNI [119] – predicting Alzheimer's disease (n = 61) versus control (n = 132) status. The results are summarized in the Table A.35.

Supplementary Figures

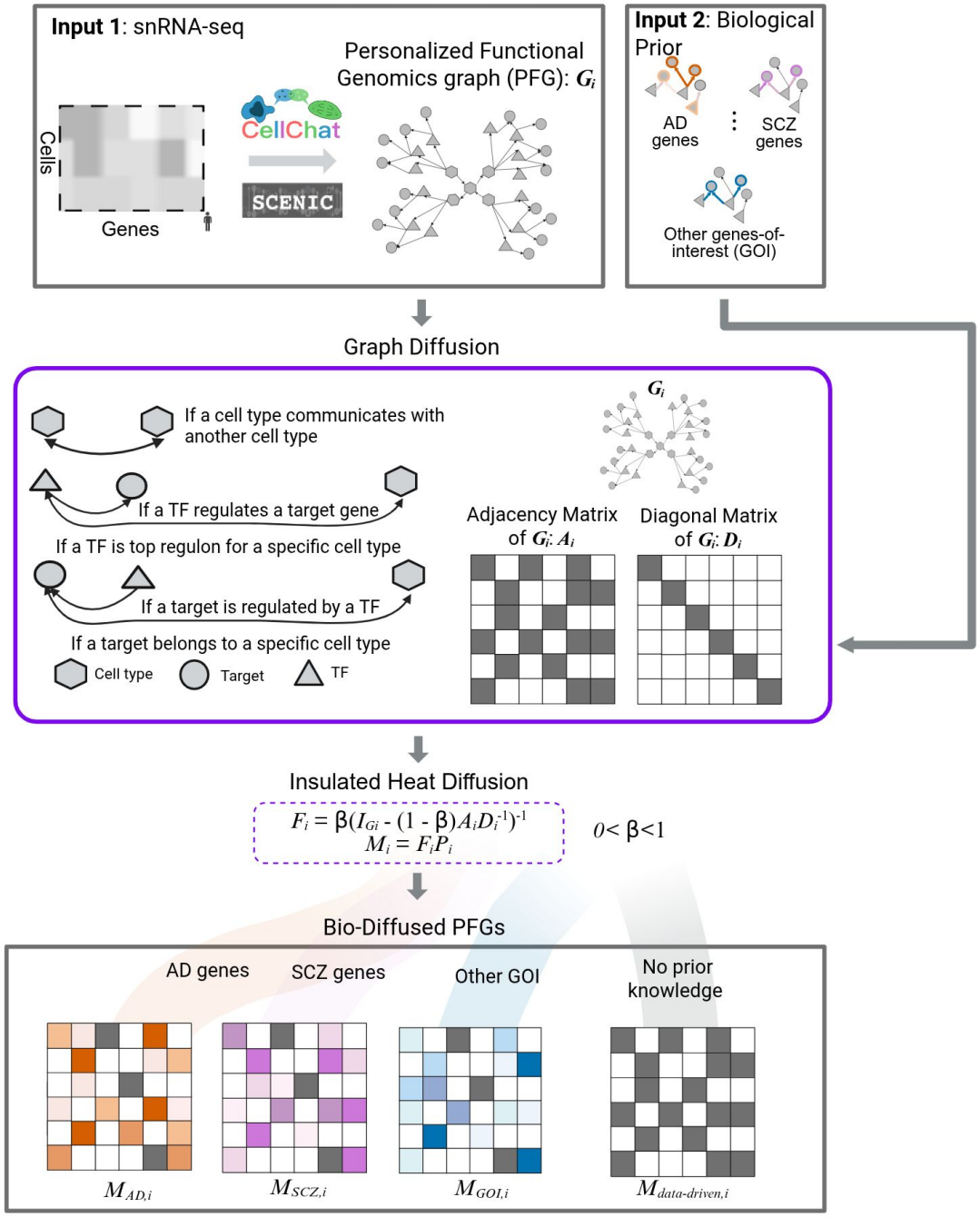

Figure A.24: **Flowchart for constructing bio-diffused PFGs.** Bio-diffused PFGs are constructed from two inputs: snRNA-seq of donors and biological genes-of-interest (GOI). First, personalized functional genomics graphs are built for each donor from their snRNA-seq using tools like CellChat and Scenic (Supplementary Note 1). Each PFG has three node types: cell types, transcription factors (TFs), and target genes (TGs), connected by directed edges. Each edge captures distinct regulatory relations, e.g., cell type interactions and TF→TG regulation. We then compute bio-diffused PFGs using insulated heat diffusion for each donor via their adjacency and diagonal matrices. In particular, the resulting bio-diffused PFGs $M^{\mathrm{AD},i}$, $M^{\mathrm{SCZ},i}$, and $M^{\mathrm{GOI},i}$ correspond to gene sets from known AD, SCZ, and other GOI for donor i. $M_{\text{data-driven}}$ is simply a matrix of ones.

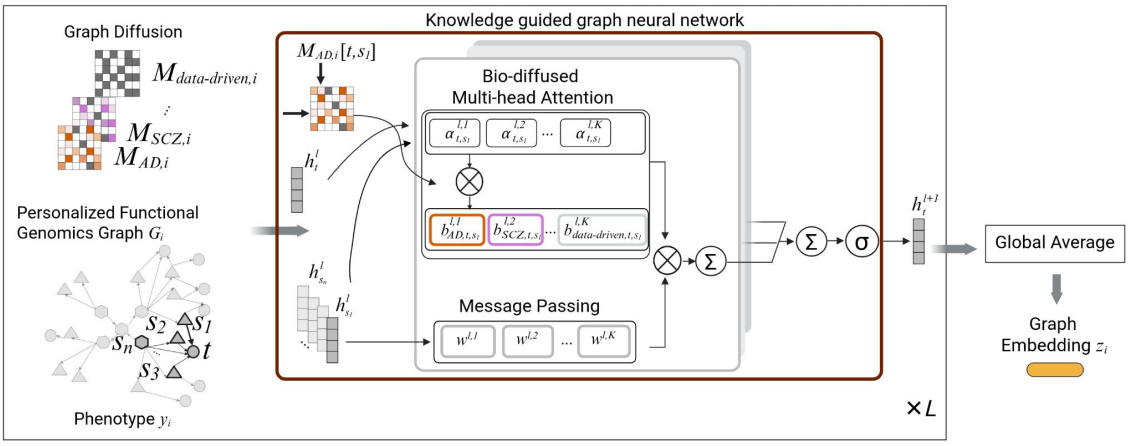

Figure A.25: **Architecture of knowledge-guided graph neural network (KG-GNN).** Given a PFG G_i, the inputs to the KG-GNN model include its bio-diffused PFGs ($M^{\text{AD},i}$, $M^{\text{SCZ},i}$, …, $M_{\text{data-driven},i}$) and node features h. The KG-GNN uses self-attention to compute scores between a node t and its neighbors s_j, incorporates priors b^{AD}_{t,s_j}, b^{SCZ}_{t,s_j}, …, $b^{\text{data-driven}}_{t,s_j}$ from the bio-diffused PFGs, normalizes them over $\mathcal{N}(t)$, and applies a softmax-weighted sum of neighbor features to update t (Supplementary Note 2).

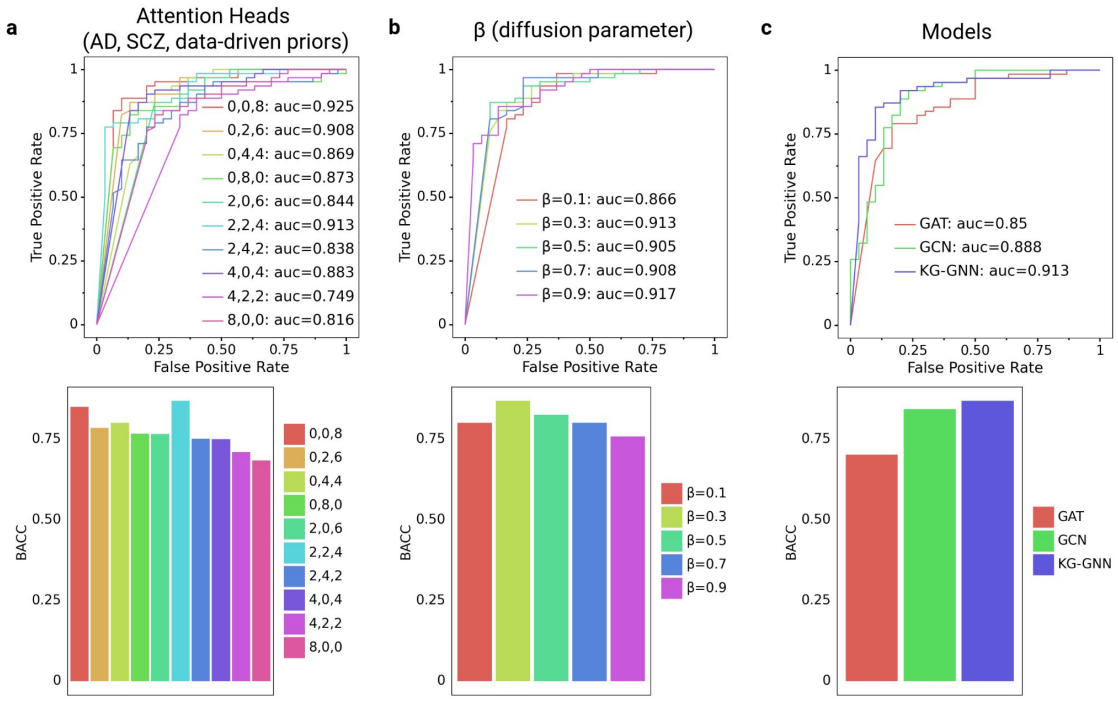

Figure A.26: **Benchmarking KG-GNN for AD vs. control classification.** ROC curves and BACC comparing (a) different KG-GNN attention-head combinations (AD, SCZ, Data-Driven), (b) diffusion parameter β, and (c) KG-GNN vs. state-of-the-art graph models: GCN, GAT (Supplementary Note 4).

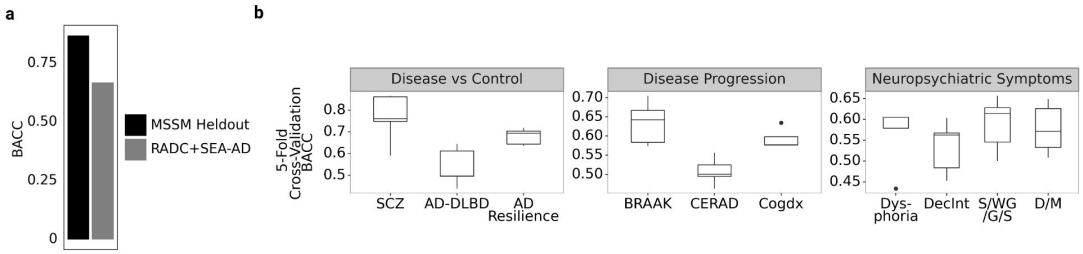

Figure A.27: **Graph embedding classification across datasets and phenotypes.** (a) BACC for KG-GNN embeddings: MSSM held-out (AD $n = 62$ vs. control $n = 30$) and RADC+SEA-AD (AD $n = 93$ vs. control $n = 68$) (Extended Fig. 1c). (b) Five-fold CV BACC across phenotype contrasts (Extended Fig. 1c): DecInt = Anhedonia, S/WG/G/S = Sleep/Weight Gain/Guilt/Suicide, D/M = Depression/Mood.

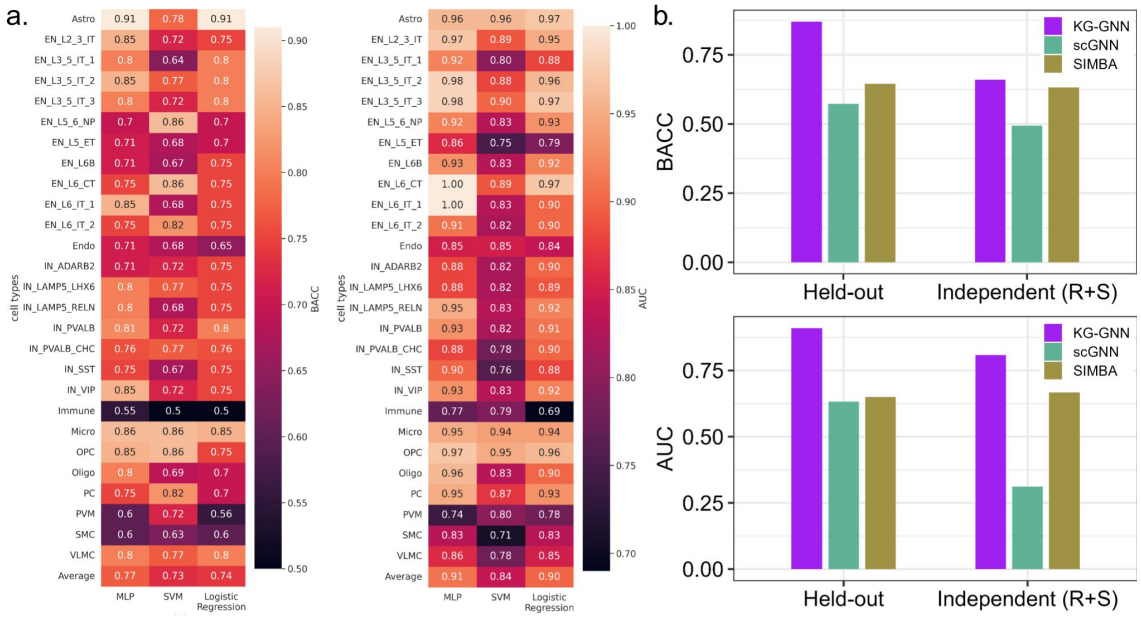

Figure A.28: **Benchmarking donor-level AD vs. control classifiers.** (a) MLP, SVM, LR on average cell-type expression: AUC and BACC (mean AUC = 0.84, range 0.96 [Astro] to 0.71 [Immune]). (b) State-of-the-art comparisons: scGNN [7], SIMBA [8].

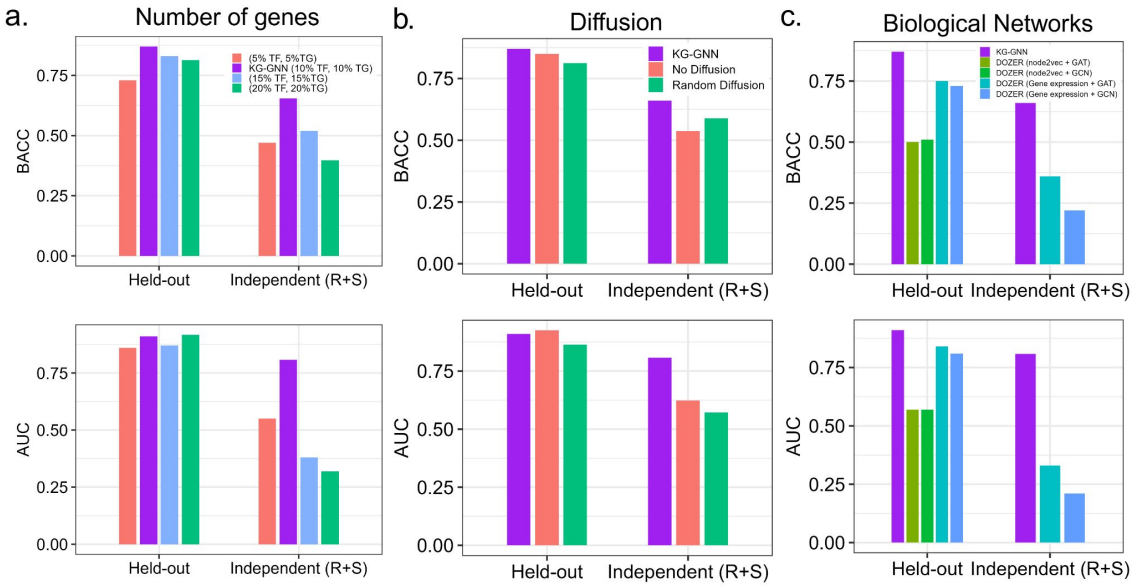

Figure A.29: **Component-wise KG-GNN evaluation for AD classification.** (a) Gene subset: top 10% TFs and top 10% TGs per TF. (b) Diffusion effect: knowledge-guided vs. random vs. none. (c) Network features: Dozer [9] co-expression, node2vec vs. expression, and GAT vs. GCN classifiers.

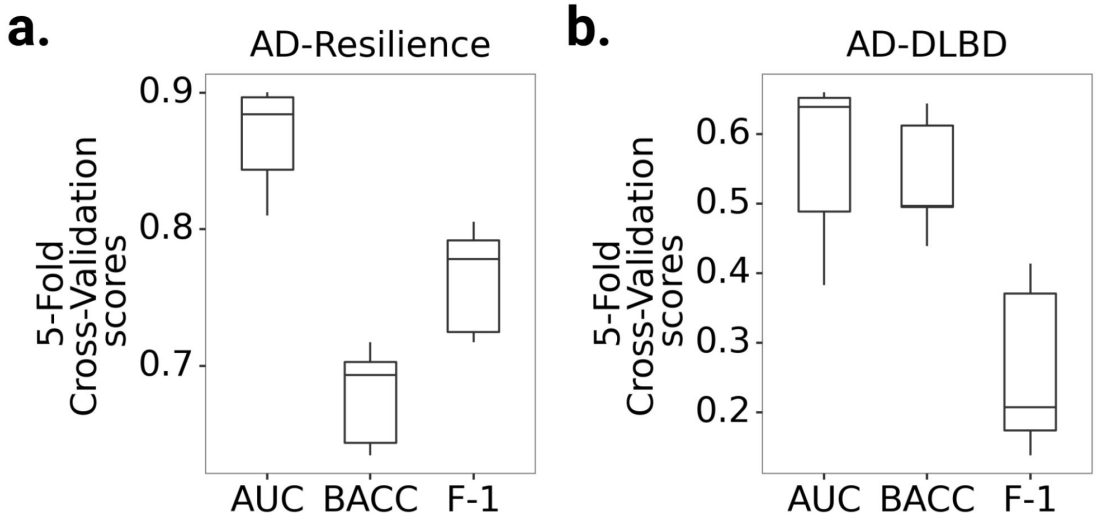

Figure A.30: **Classification of AD-related phenotypes.** (a) AD-Resilience: AUC = 0.87, BACC = 0.68, F1 = 0.76. (b) AD-DLBD: AUC = 0.56, BACC = 0.53, F1 = 0.26.

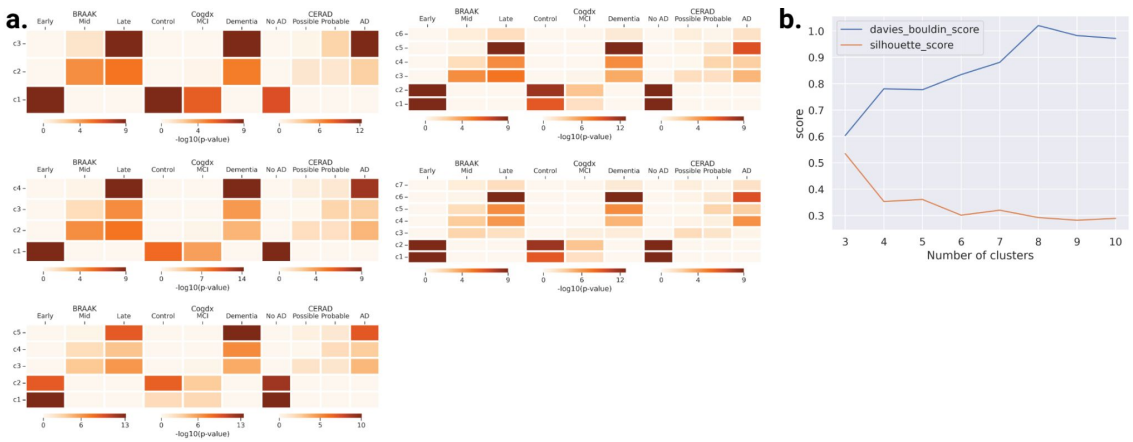

Figure A.31: **Clustering donor embeddings across AD phenotypes.** (a) Heatmaps of phenotype enrichment for cluster counts 3–7 (BRAAK, CERAD, CDR) via hypergeometric p-values. (b) Davies–Bouldin vs. Silhouette scores for 3–7 clusters: compactness vs. separation metrics.

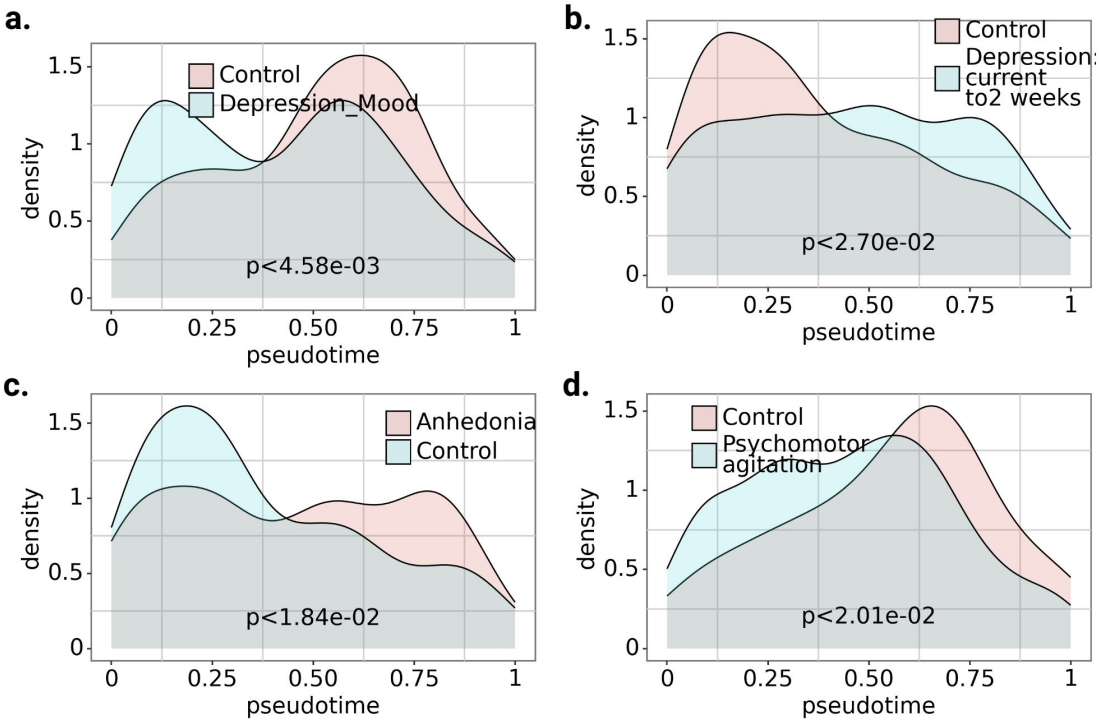

Figure A.32: **NPS distribution across donors.** Density plots of neuropsychiatric symptoms: (a) Depression: Mood ($p < 1.85 \times 10^{-2}$), (b) Depression (current 2 weeks) ($p < 2.70 \times 10^{-2}$), (c) DecInt: Anhedonia ($p < 1.84 \times 10^{-2}$), (d) Psychomotor agitation ($p < 2.01 \times 10^{-2}$).

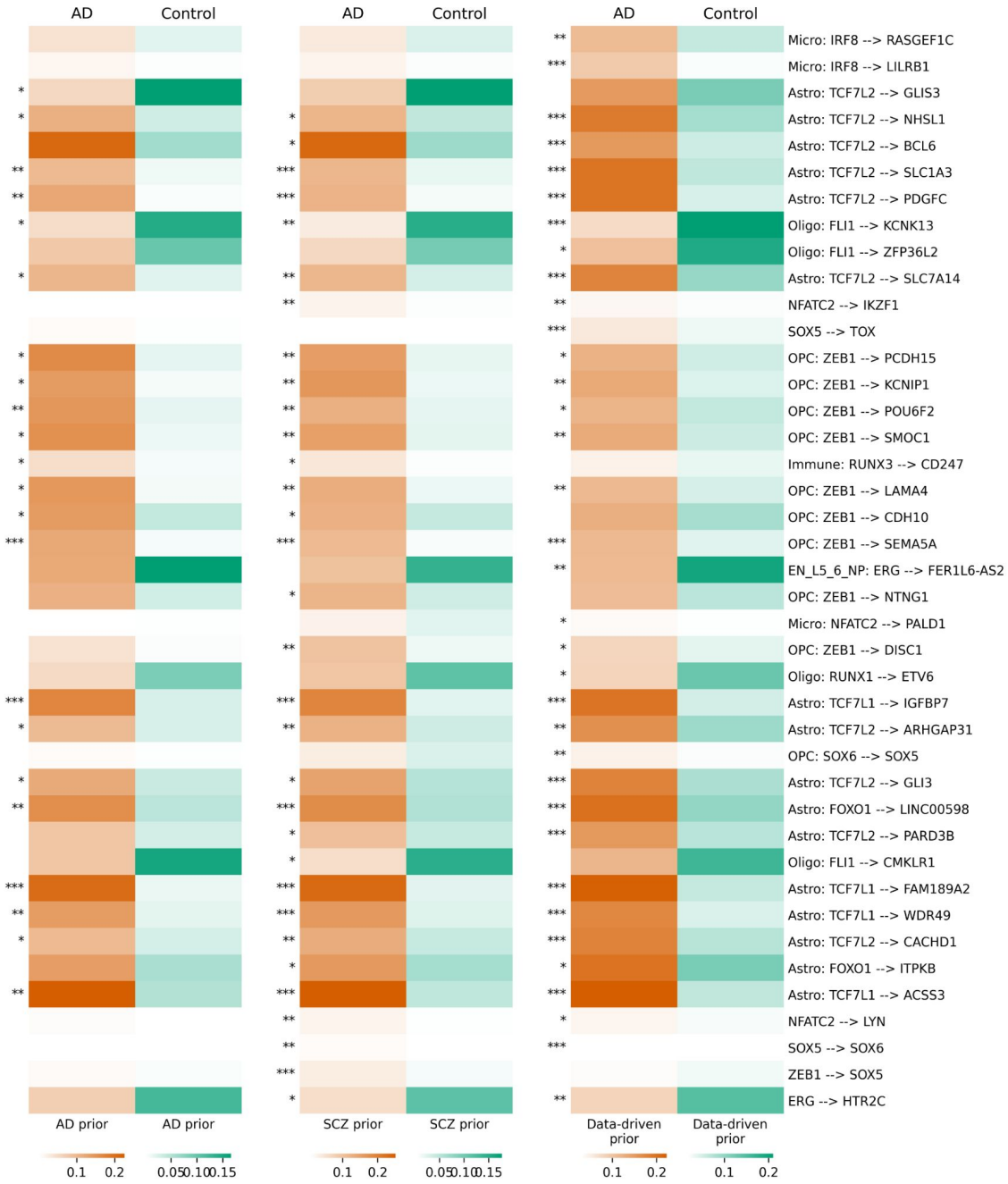

Figure A.33: **TF-TG link importance across priors.** Importance scores for TF→TG edges in AD vs. control donors under AD, SCZ, and data-driven priors. Significance: * $=p < 0.05$, ** $=p < 0.01$, *** $=p < 0.001$.

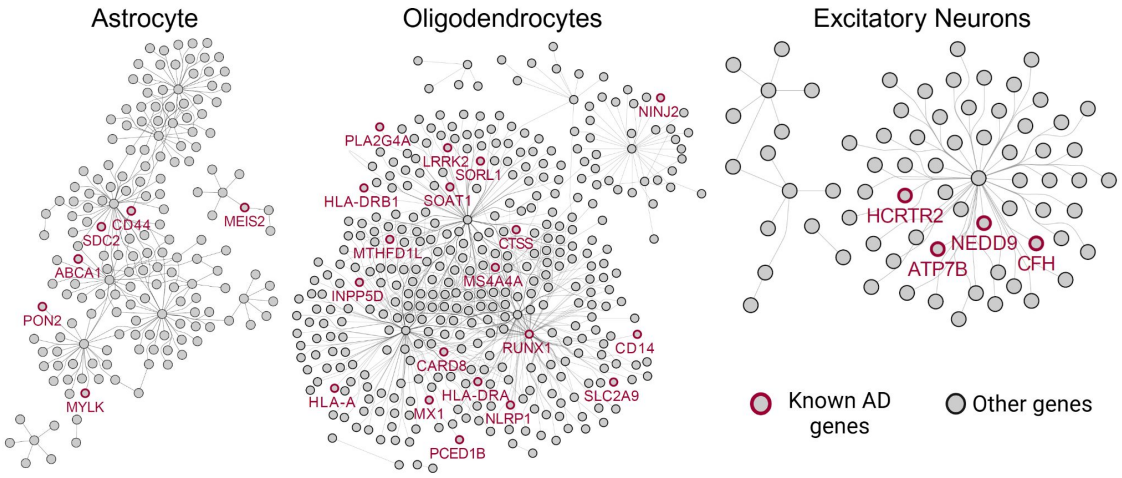

Figure A.34: **Prioritized subnetworks in main cell clusters.** For the three cell-type clusters from Fig. 4e: ellipses are genes; known AD genes have orange borders; other gene names hidden for clarity.

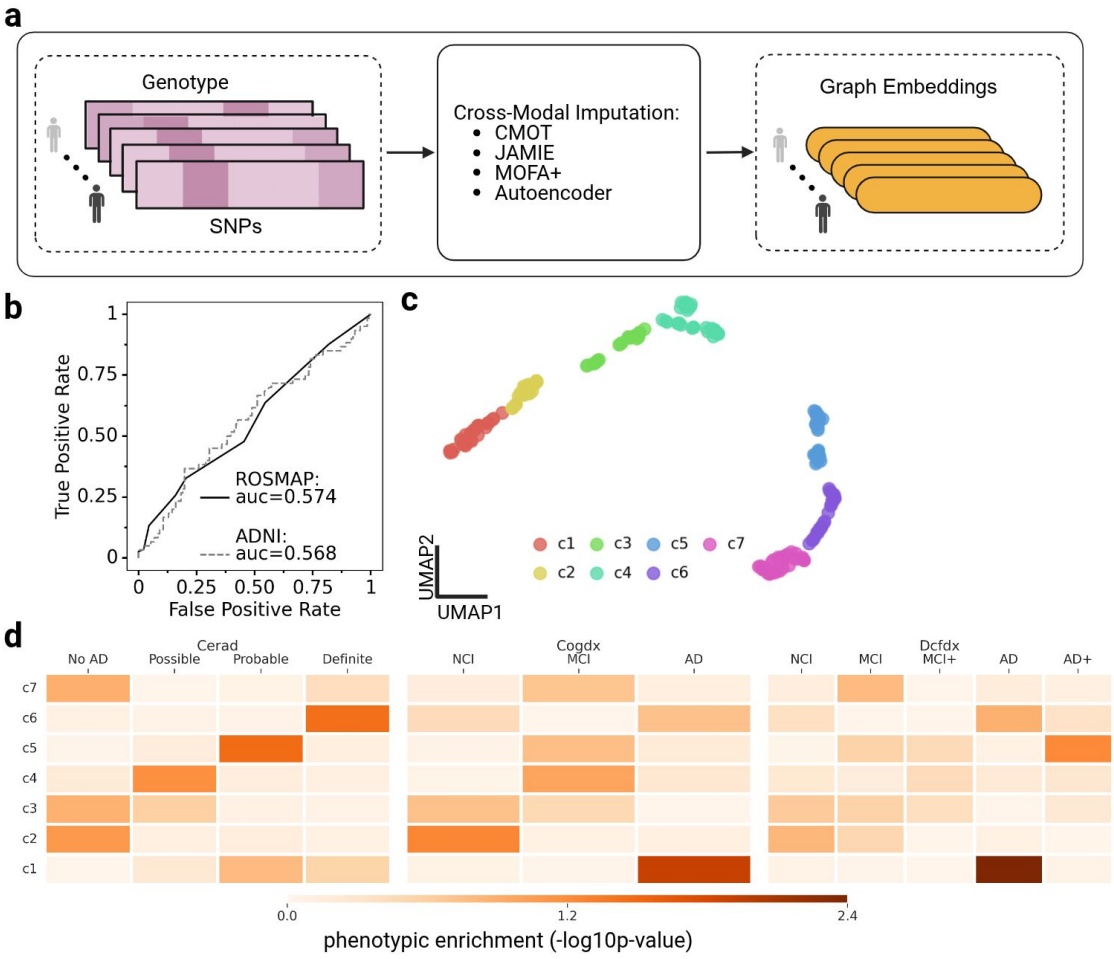

Figure A.35: **Independent validation of graph-embedding imputation.** (a) Cross-modal imputation methods: CMOT[4], JAMIE[29], MOFA+[9], Autoencoder[10]. (b) ROC for imputed embeddings: ROSMAP BRAAK early ($n = 88$) vs. late ($n = 113$), ADNI AD ($n = 61$) vs. control ($n = 132$). (c) UMAP of ROSMAP embeddings with unsupervised clustering. (d) Heatmaps of cluster phenotype enrichments (CERAD, Cogdx, Dcfdx) via hypergeometric test.

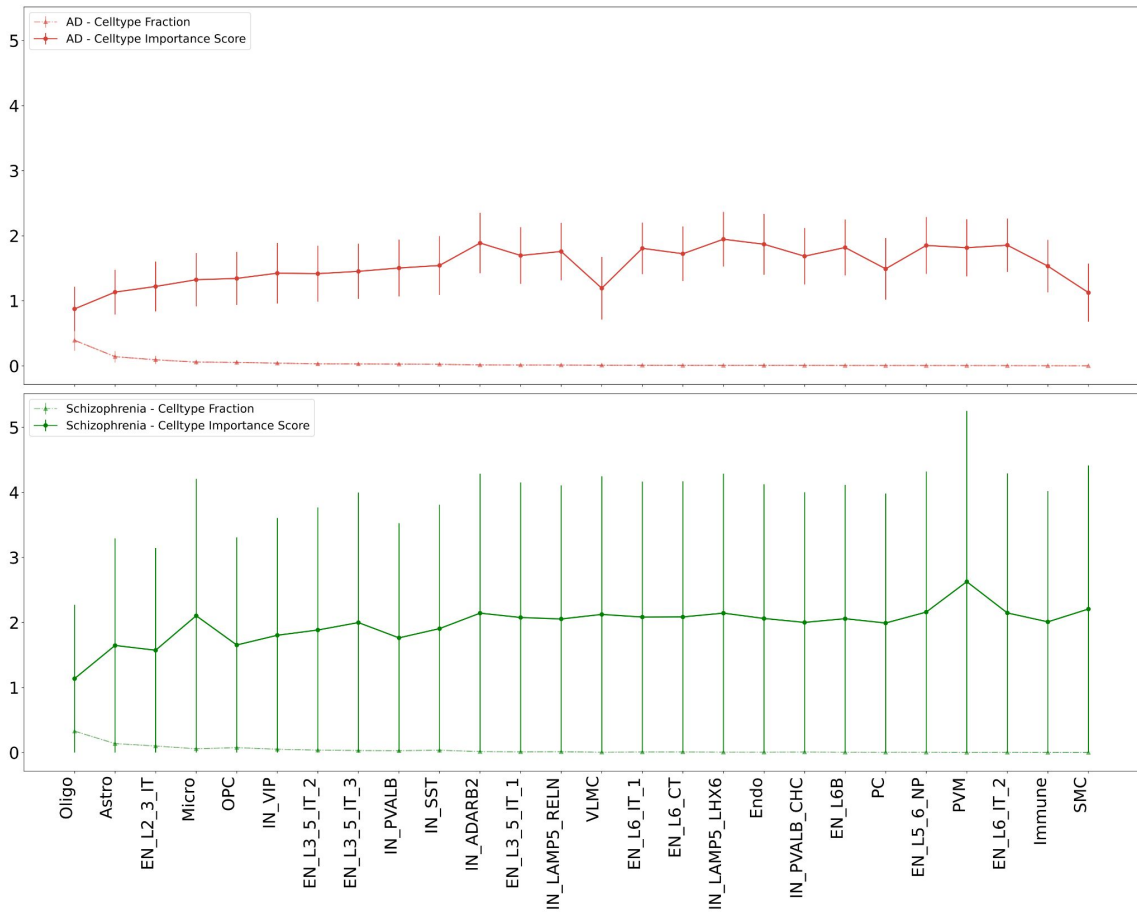

Figure A.36: **Cell-type importance scores for AD and SCZ donors.** Circles = importance scores; triangles = cell fractions per cell type.

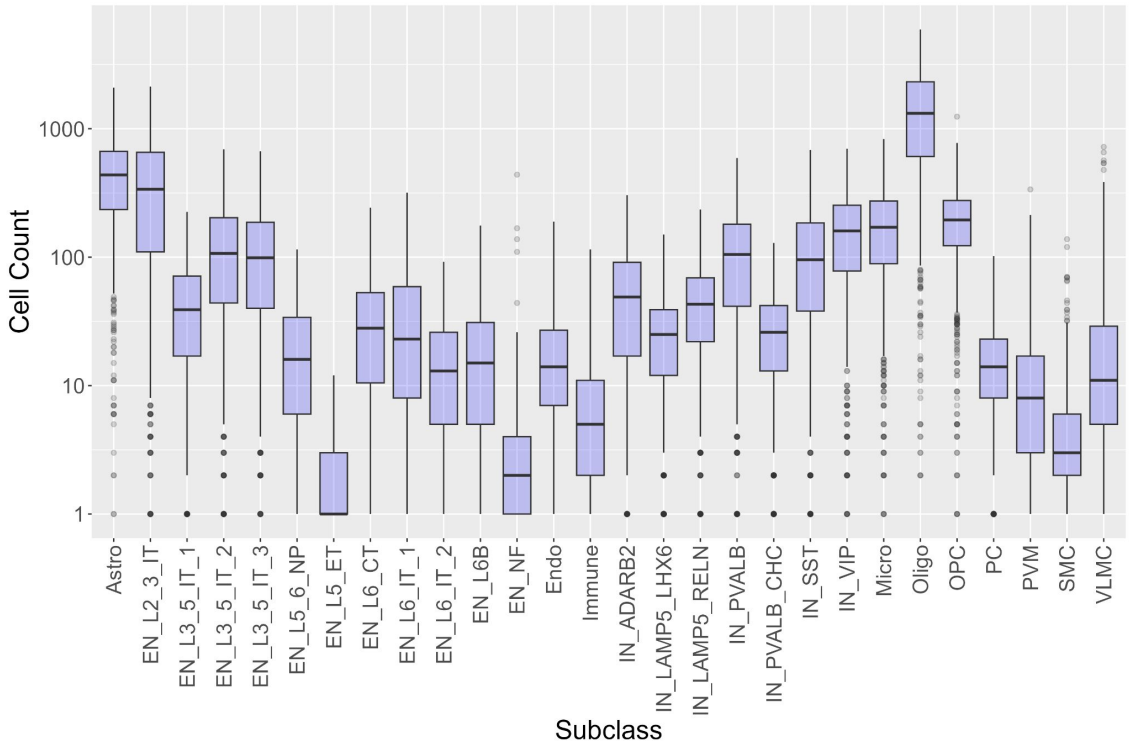

Figure A.37: **Cell counts per cell type across all donors.** The x-axis indicates the cell types, and the y-axis is the cell counts in the log scale for each cell type.

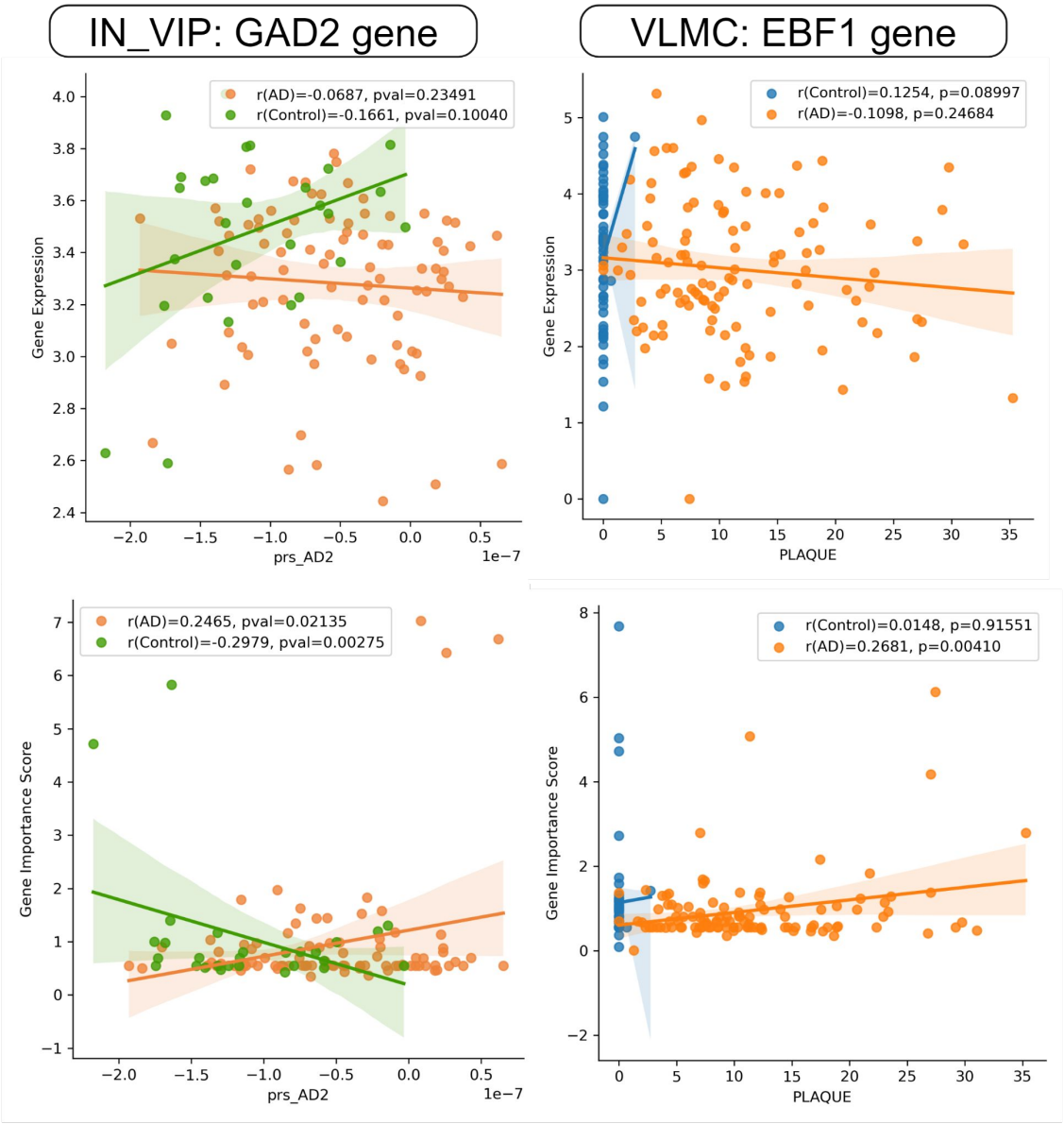

Figure A.38: **Correlation comparison of gene importance score and gene expression of some select genes.**

Supplementary Tables

Table A.34: **KG-GNN final hyperparameters, training, and model details**

Hyperparameters	
Learning rate	1×10^{-4}
Batch size	5
Diffusion parameter (β)	0.3
Dropout	0.6
Number of subgraphs	3
Number of attention heads	8
Training	
Epochs	23
Memory (MB) per epoch	36,406
Total runtime (s)	1,096.56
Model Architecture	
GATConv → BatchNorm → LayerNorm	4 layers (2048, 1024, 512, 256)
Linear → ReLU → Dropout	2 layers (128, 64)
Linear → softmax	1 layer

Table A.35: **Imputation performance of graph embedding using genotype data**

Method	AD vs. control (ADNI)		Early vs. Late BRAAK stage (ROSMAP)	
	BACC	AUC	BACC	AUC
CMOT	0.58	0.57	0.56	0.57
JAMIE	0.53	0.55	0.50	0.54
MOFA+	0.49	0.55	0.48	0.54
AutoEncoder	0.54	0.57	0.54	0.53

Table A.36: **Correlation comparison of gene importance score vs. gene expression for different phenotypes.**

Phenotype	Clinical phenotype	mean_corr_importance_score	mean_corr_gene_expression	T-test p-value
AD_c15x	Age	0.0978	0.0644	2.01×10^{-7}
AD_c15x	CDR	0.1144	0.0819	6.29×10^{-8}
AD_c15x	Cog_Tau_Resilience	0.1274	0.0550	1.85×10^{-30}
AD_c15x	PLAQUE	0.0773	0.0590	0.0165
AD_c15x	prs_AD2	0.1237	0.0656	7.78×10^{-12}
AD_resiliency	Age	0.2008	0.0615	1.22×10^{-7}
AD_resiliency	CDR	0.1657	0.0923	1.25×10^{-21}
AD_resiliency	Cog_Tau_Resilience	0.1811	0.0893	7.53×10^{-21}
AD_resiliency	PLAQUE	0.1547	0.2037	0.0311
AD_resiliency	prs_AD2	0.2095	0.1103	6.58×10^{-31}
EarlyInsom	Age	0.2071	0.1187	3.00×10^{-12}
EarlyInsom	CDR	0.1960	0.1044	6.57×10^{-14}
EarlyInsom	Cog_Tau_Resilience	0.2173	0.1473	1.57×10^{-5}
EarlyInsom	PLAQUE	0.2085	0.1360	1.29×10^{-7}
EarlyInsom	prs_AD2	0.2831	0.1364	1.79×10^{-20}
LateInsom	Age	0.2107	0.1123	7.95×10^{-8}
LateInsom	CDR	0.1735	0.1111	3.09×10^{-5}
LateInsom	Cog_Tau_Resilience	0.2229	0.0988	5.76×10^{-10}
LateInsom	PLAQUE	0.2145	0.1093	3.03×10^{-9}
LateInsom	prs_AD2	0.3454	0.1688	3.03×10^{-10}
MSSM_SWGS	Age	0.1490	0.0924	3.78×10^{-13}
MSSM_SWGS	CDR	0.1521	0.0998	1.26×10^{-12}
MSSM_SWGS	Cog_Tau_Resilience	0.1678	0.0856	9.56×10^{-25}
MSSM_SWGS	PLAQUE	0.1579	0.0936	1.91×10^{-18}
MSSM_SWGS	prs_AD2	0.2139	0.0999	4.43×10^{-31}
MidInsom	Age	0.2440	0.1370	3.95×10^{-7}
MidInsom	CDR	0.2275	0.0963	7.95×10^{-12}
MidInsom	Cog_Tau_Resilience	0.2487	0.1254	5.23×10^{-8}
MidInsom	PLAQUE	0.2248	0.1127	2.30×10^{-8}
MidInsom	prs_AD2	0.2955	0.1375	7.19×10^{-15}
SCZ_c07x	Age	0.1221	0.0802	3.60×10^{-8}
SCZ_c07x	CDR	0.1266	0.0865	1.18×10^{-14}
SCZ_c07x	Cog_Tau_Resilience	0.1342	0.0586	3.34×10^{-48}
SCZ_c07x	PLAQUE	0.1020	0.0892	0.1738
SCZ_c07x	prs_AD2	0.1443	0.0740	1.02×10^{-19}

Bibliography

[1] Manifold learning theory and applications.

[2] Quadri Adewale, Ahmed F Khan, David A Bennett, and Yasser Iturria-Medina. Single-nucleus rna velocity reveals critical synaptic and cell-cycle dysregulations in neuropathologically confirmed alzheimer's disease. *Scientific Reports*, 14(1):7269, 2024.

[3] Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, et al. Scenic: single-cell regulatory network inference and clustering. *Nature methods*, 14(11):1083–1086, 2017.

[4] Sayali Anil Alatkar and Daifeng Wang. Cmot: cross-modality optimal transport for multimodal inference. *Genome Biology*, 24(1):163, 2023.

[5] Sayali Anil Alatkar and Daifeng Wang. Artemis integrates autoencoders and schrödinger bridges to predict continuous dynamics of gene expression, cell population, and perturbation from time-series single-cell data. *Bioinformatics*, $41(\text{Supplement}_1) : i189 - -i197, 072025$.

[6] David Alvarez-Melis and Tommi S. Jaakkola. Gromov-wasserstein alignment of word embedding spaces.

[7] Ashlyn G Anderson, Brianne B Rogers, Jacob M Loupe, Ivan Rodriguez-Nunez, Sydney C Roberts, Lauren M White, J Nicholas Brazell, William E Bunney, Blynn G Bunney, Stanley J Watson, et al. Single nucleus multiomics identifies zeb1 and mafb as candidate regulators of alzheimer's disease-specific cis-regulatory elements. *Cell Genomics*, 3(3), 2023.

[8] Ricard Argelaguet, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C. Marioni, and Oliver Stegle. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. 21(1):111.

[9] Ricard Argelaguet, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C Marioni, and Oliver Stegle. Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome biology*, 21:1–17, 2020.

[10] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders. *Machine learning for data science handbook: data mining and knowledge discovery handbook*, pages 353–374, 2023.

[11] Ada Bernaus, Sandra Blanco, and Ana Sevilla. Glia crosstalk in neuroinflammatory diseases. *Frontiers in cellular neuroscience*, 14:209, 2020.

[12] Espen Bernton, Jeremy Heng, Arnaud Doucet, and Pierre E Jacob. Schr\" odinger bridge samplers. *arXiv preprint arXiv:1912.13170*, 2019.

[13] Lukas Biewald. Experiment tracking with weights and biases, 2020. Software available from wandb.com.

[14] Danielle Bittencourt, Dai-Ying Wu, Kwang Won Jeong, Daniel S. Gerke, Laurie Herviou, Irina Ianculescu, Rajas Chodankar, Kimberly D. Siegmund, and Michael R. Stallcup. G9a functions as a molecular scaffold for assembly of transcriptional coactivators on a subset of glucocorticoid receptor target genes. 109(48):19673–19678.

[15] Patricia A Boyle, Tianhao Wang, Lei Yu, Robert S Wilson, Robert Dawe, Konstantinos Arfanakis, Julie A Schneider, and David A Bennett. To what degree is late life cognitive decline driven by age-related neuropathologies? *Brain*, 144(7):2166–2175, 2021.

[16] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs. 2018.

[17] Junyue Cao, Darren A. Cusanovich, Vijay Ramani, Delasa Aghamirzaie, Hannah A. Pliner, Andrew J. Hill, Riza M. Daza, Jose L. McFaline-Figueroa, Jonathan S. Packer, Lena Christiansen, Frank J. Steemers, Andrew C. Adey, Cole Trapnell, and Jay Shendure. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. 361(6409):1380–1385.

[18] Kai Cao, Xiangqi Bai, Yiguang Hong, and Lin Wan. Unsupervised topological alignment for single-cell multi-omics integration. 36:i48–i56.

[19] Kai Cao, Yiguang Hong, and Lin Wan. Manifold alignment for heterogeneous single-cell multi-omics data integration using pamona. 38(1):211–219.

[20] Zhi-Jie Cao and Ge Gao. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 40(10):1458–1466, 2022.

[21] Minerva M Carrasquillo, Fanggeng Zou, V Shane Pankratz, Samantha L Wilcox, Li Ma, Louise P Walker, Samuel G Younkin, Curtis S Younkin, Linda H Younkin, Gina D Bisceglio, et al. Genetic variation in pcdh11x is associated with susceptibility to late-onset alzheimer's disease. *Nature genetics*, 41(2):192–198, 2009.

[22] Cayley. On monge's "mémoire sur la théorie des déblais et des remblais.". s1-14(1):139–143.

[23] Pramod Bharadwaj Chandrashekar, Sayali Alatkar, Jiebiao Wang, Gabriel E Hoffman, Chenfeng He, Ting Jin, Saniya Khullar, Jaroslav Bendl, John F Fullard, Panos Roussos, et al. Deepgami: deep biologically guided auxiliary learning for multimodal integration and imputation to improve genotype–phenotype prediction. *Genome Medicine*, 15(1):88, 2023.

[24] Pramod Bharadwaj Chandrashekar, Sayali Anil Alatkar, Noah Cohen Kalafut, Ting Jin, Chirag Gupta, Ryan Burzak, Xiang Huang, Shuang Liu, Athan Z. Li, PsychAD Consortium, Kiran Girdhar, Georgios Voloudakis, Gabriel E. Hoffman, Jaroslav Bendl, John F. Fullard, Donghoon Lee, Panos Roussos, and Daifeng Wang. Personalized single-cell transcriptomics reveals molecular diversity in alzheimer's disease. *medRxiv*, 2024.

[25] Huidong Chen, Jayoung Ryu, Michael E Vinyard, Adam Lerer, and Luca Pinello. Simba: single-cell embedding along with features. *Nature Methods*, 21(6):1003–1013, 2024.

[26] Song Chen, Blue B. Lake, and Kun Zhang. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. 37(12):1452–1457.

[27] Tianrong Chen, Guan-Horng Liu, and Evangelos A Theodorou. Likelihood training of schrödinger bridge using forward-backward sdes theory. In *International Conference on Learning Representations*, 2022.

[28] Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and kantorovich formulations. 274(11):3090–3123.

[29] Noah Cohen Kalafut, Xiang Huang, and Daifeng Wang. Joint variational autoencoders for multimodal imputation and embedding. *Nature machine intelligence*, 5(6):631–642, 2023.

[30] Micaela E Consens, Yuxiao Chen, Vilas Menon, Yanling Wang, Julie A Schneider, Philip L De Jager, David A Bennett, Shreejoy J Tripathy, and Daniel Felsky. Bulk and single-nucleus transcriptomics highlight intra-telencephalic and somatostatin neurons in alzheimer's disease. *Frontiers in Molecular Neuroscience*, 15:903175, 2022.

[31] David P Cook and Barbara C Vanderhyden. Context specificity of the emt transcriptional response. *Nature communications*, 11(1):2142, 2020.

[32] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *CoRR*, abs/1507.00504, 2015.

[33] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[34] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances, 2013.

[35] Marco Cuturi, Laetitia Meng-Papaxanthos, Yingtao Tian, Charlotte Bunne, Geoff Davis, and Olivier Teboul. Optimal transport tools (ott): A jax toolbox for all things wasserstein. *arXiv preprint arXiv:2201.12324*, 2022.

[36] Louis De Beaumont, Sandra Pelleieux, Louise Lamarre-Théroux, Doris Dea, Judes Poirier, and Alzheimer's Disease Cooperative Study. Butyrylcholinesterase k and apolipoprotein e-4 reduce the age of onset of alzheimer's disease, accelerate cognitive decline, and modulate donepezil response in mild cognitively impaired subjects. *Journal of Alzheimer's Disease*, 54(3):913–922, 2016.

[37] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021.

[38] Philip L De Jager, Yiyi Ma, Cristin McCabe, Jishu Xu, Badri N Vardarajan, Daniel Felsky, Hans-Ulrich Klein, Charles C White, Mette A Peters, Ben Lodgson, et al. A multi-omic atlas of the human frontal cortex for aging and alzheimer's disease research. *Scientific data*, 5(1):1–13, 2018.

[39] Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh. SCOT: Single-cell multi-omics alignment with optimal transport. 29(1):3–18.

[40] Pınar Demetçi, Rebecca Santorella, Björn Sandstede, and Ritambhara Singh. Unsupervised integration of single-cell multi-omics datasets with disproportionate cell-type representation. In Itsik Pe'er, editor, *Research in Computational Molecular Biology*, volume 13278, pages 3–19. Springer International Publishing. Series Title: Lecture Notes in Computer Science.

[41] Constanze Depp, Ting Sun, Andrew Octavian Sasmita, Lena Spieth, Stefan A Berghoff, Taisiia Nazarenko, Katharina Overhoff, Agnes A Steixner-Kumar, Swati Subramanian, Sahab Arinrad, et al. Myelin dysfunction drives amyloid-β deposition in models of alzheimer's disease. *Nature*, 618(7964):349–357, 2023.

[42] Maria A. Dimitriu, Irina Lazar-Contes, Martin Roszkowski, and Isabelle M. Mansuy. Single-cell multiomics techniques: From conception to applications. 10:854317.

[43] Jinzhuang Dou, Shaoheng Liang, Vakul Mohanty, Qi Miao, Yuefan Huang, Qingnan Liang, Xuesen Cheng, Sangbae Kim, Jongsu Choi, Yumei Li, et al. Bi-order multimodal integration of single-cell data. *Genome biology*, 23(1):112, 2022.

[44] Prashant S Emani, Jason J Liu, Declan Clarke, Matthew Jensen, Jonathan Warrell, Chirag Gupta, Ran Meng, Che Yu Lee, Siwei Xu, Cagatay Dursun, et al. Single-cell genomics and regulatory networks for 388 human brains. *Science*, 384(6698):eadi5199, 2024.

[45] Pedregosa et al. Scikit-learn: Machine learning in python.

[46] Jeffrey A Farrell, Yiqun Wang, Samantha J Riesenfeld, Karthik Shekhar, Aviv Regev, and Alexander F Schier. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, 360(6392):eaar3131, 2018.

[47] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*, 2019.

[48] Robert Fortet. Résolution d'un système d'équations de m. schrödinger. *Journal de Mathématiques Pures et Appliquées*, 19(1-4):83–105, 1940.

[49] Anna S Fröhlich, Nathalie Gerstner, Miriam Gagliardi, Maik Ködel, Natan Yusupov, Natalie Matosin, Darina Czamara, Susann Sauer, Simone Roeh, Vanessa Murek, et al. Single-nucleus transcriptomic profiling of human orbitofrontal cortex reveals convergent effects of aging and psychiatric disease. *Nature Neuroscience*, 27(10):2021–2032, 2024.

[50] Lukas Frontzkowski, Michael Ewers, Matthias Brendel, Davina Biel, Rik Ossenkoppele, Paul Hager, Anna Steward, Anna Dewenter, Sebastian Römer, Anna Rubinski, et al. Earlier alzheimer's disease onset is associated with tau pathology in brain hub regions and facilitated tau spreading. *Nature communications*, 13(1):4899, 2022.

[51] John F Fullard, Prashant Nm, Donghoon Lee, Deepika Mathur, Karen Therrien, Aram Hong, Clara Casey, Zhiping Shao, Marcela Alvia, Stathis Argyriou, et al. Population-scale cross-disorder atlas of the human prefrontal cortex at single-cell resolution. *Scientific Data*, 12(1):1–13, 2025.

[52] Mariano I Gabitto, Kyle J Travaglini, Victoria M Rachleff, Eitan S Kaplan, Brian Long, Jeanelle Ariza, Yi Ding, Joseph T Mahoney, Nick Dee, Jeff Goldy, Erica J Melief, Krissy Brouner, Jazmin Campos, John Campos, Ambrose J Carr, Tamara Casper, Rushil Chakrabarty, Michael Clark, Jonah Cool, Nasmil J Valera Cuevas, Rachel Dalley, Martin Darvas, Song-Lin Ding, Tim Dolbeare, Christine L Mac Donald, Tom Egdorf, Luke Esposito, Rebecca Ferrer, Rohan Gala, Amanda Gary, Jessica Gloe, Nathan Guilford, Junitta Guzman, Daniel Hirschstein, Windy Ho, Tim Jarksy, Nelson Johansen, Brian E Kalmbach, Lisa M Keene, Sarah Khawand, Mitch Kilgore, Amanda Kirkland, Michael Kunst, Brian R Lee, Jocelin Malone, Zoe Maltzer, Naomi Martin, Rachel McCue, Delissa McMillen, Emma Meyerdierks, Kelly P Meyers, Tyler Mollenkopf, Mark Montine, Amber L Nolan, Julie Nyhus, Paul A Olsen, Maiya Pacleb, Nicholas Peña, Thanh Pham, Christina Alice Pom, Nadia Postupna, Augustin Ruiz, Aimee M Schantz, Nadiya V Shapovalova, Staci A Sorensen, Brian Staats, Matt Sullivan, Susan M Sunkin, Carol Thompson, Michael Tieu, Jonathan Ting, Amy Torkelson, Tracy Tran, Ming-Qiang Wang, Jack Waters, Angela M Wilson, David Haynor, Nicole Gatto, Suman Jayadev, Shoaib Mufti, Lydia Ng, Shubhabrata Mukherjee, Paul K Crane, Caitlin S Latimer, Boaz P Levi, Kimberly Smith, Jennie L Close, Jeremy A Miller, Rebecca D Hodge, Eric B Larson,

Thomas J Grabowski, Michael Hawrylycz, C Dirk Keene, and Ed S Lein. Integrated multimodal cell atlas of alzheimer's disease. May 2023.

[53] Julia Gamache, Daniel Gingerich, E Keats Shwab, Julio Barrera, Melanie E Garrett, Cordelia Hume, Gregory E Crawford, Allison E Ashley-Koch, and Ornit Chiba-Falek. Integrative single-nucleus multi-omics analysis prioritizes candidate cis and trans regulatory networks and their target genes in alzheimer's disease brains. *Cell & Bioscience*, 13(1):185, 2023.

[54] Adam Gayoso, Zoë Steier, Romain Lopez, Jeffrey Regier, Kristopher L. Nazor, Aaron Streets, and Nir Yosef. Joint probabilistic modeling of single-cell multi-omic data with totalVI. 18(3):272–282.

[55] Adam Gayoso, Philipp Weiler, Mohammad Lotfollahi, Dominik Klein, Justin Hong, Aaron Streets, Fabian J Theis, and Nir Yosef. Deep generative modeling of transcriptional dynamics for rna velocity analysis in single cells. *Nature methods*, 21(1):50–59, 2024.

[56] Gilad Sahar Green, Masashi Fujita, Hyun-Sik Yang, Mariko Taga, Anael Cain, Cristin McCabe, Natacha Comandante-Lou, Charles C White, Anna K Schmidtner, Lu Zeng, et al. Cellular communities reveal trajectories of brain ageing and alzheimer's disease. *Nature*, 633(8030):634–645, 2024.

[57] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

[58] Alexandra Grubman, Gabriel Chew, John F Ouyang, Guizhi Sun, Xin Yi Choo, Catriona McLean, Rebecca K Simmons, Sam Buckberry, Dulce B Vargas-Landin, Daniel Poppe, et al. A single-cell atlas of entorhinal cortex from individuals with alzheimer's disease reveals cell-type-specific gene expression regulation. *Nature neuroscience*, 22(12):2087–2097, 2019.

[59] Chirag Gupta, Jielin Xu, Ting Jin, Saniya Khullar, Xiaoyu Liu, Sayali Alatkar, Feixiong Cheng, and Daifeng Wang. Single-cell network biology characterizes cell type gene regulation for drug repurposing and phenotype prediction in alzheimer's disease. *PLoS Computational Biology*, 18(7):e1010287, 2022.

[60] Naomi Habib, Cristin McCabe, Sedi Medina, Miriam Varshavsky, Daniel Kitsberg, Raz Dvir-Szternfeld, Gilad Green, Danielle Dionne, Lan Nguyen, Jamie L Marshall, et al. Disease-associated astrocytes in alzheimer's disease and aging. *Nature neuroscience*, 23(6):701–706, 2020.

[61] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. 20(1):296.

[62] Laleh Haghverdi, Maren Büttner, F. Alexander Wolf, Florian Buettner, and Fabian J Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13:845–848, 2016.

[63] Laleh Haghverdi, Maren Büttner, F Alexander Wolf, Florian Buettner, and Fabian J Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature methods*, 13(10):845–848, 2016.

[64] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[65] David V Hansen, Jesse E Hanson, and Morgan Sheng. Microglia in alzheimer's disease. *Journal of Cell Biology*, 217(2):459–472, 2018.

[66] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. 184(13):3573–3587.e29.

[67] Yuhan Hao, Tim Stuart, Madeline H Kowalski, Saket Choudhary, Paul Hoffman, Austin Hartman, Avi Srivastava, Gesmira Molla, Shaista Madad, Carlos Fernandez-Granda, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature biotechnology*, 42(2):293–304, 2024.

[68] Michael Hawrylycz, Eitan S Kaplan, Kyle J Travaglini, Mariano I Gabitto, Jeremy A Miller, Lydia Ng, Jennie L Close, Rebecca D Hodge, Brian Long, Tyler Mollenkopf, et al. Sea-ad is a multimodal cellular atlas and resource for alzheimer's disease. *Nature Aging*, 4(10):1331–1334, 2024.

[69] Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3, 2017.

[70] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[71] Katherine A Hoadley, Christina Yau, Toshinori Hinoue, Denise M Wolf, Alexander J Lazar, Esther Drill, Ronglai Shen, Alison M Taylor, Andrew D Cherniack, Vésteinn Thorsson, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2):291–304, 2018.

[72] Lars Holdijk, Yuanqi Du, Priyank Jaini, Ferry Hooft, Bernd Ensing, and Max Welling. Path integral stochastic optimal control for sampling transition paths. In *ICML 2022 2nd AI for Science Workshop*, 2022.

[73] C Howald. Combining RT-PCR-seq and RNA-seq to catalog all genic elements encoded in the human genome. *Genome Res*, 22:1698–1710, 2012.

[74] Jian Huang, Yuling Jiao, Lican Kang, Xu Liao, Jin Liu, and Yanyan Liu. Schrödinger-föllmer sampler: sampling without ergodicity. *arXiv preprint arXiv:2106.10880*, 1, 2021.

[75] Jiawei Huang, Jie Sheng, and Daifeng Wang. Manifold learning analysis suggests strategies to align single-cell multimodal data of neuronal electrophysiology and transcriptomics. 4(1):1308.

[76] Xiang Huang, Noah Cohen Kalafut, Sayali Alatkar, Athan Z Li, Qiping Dong, Qiang Chang, and Daifeng Wang. Neurotd: A time-frequency based multimodal learning approach to analyze time delays in neural activities. *bioRxiv*, pages 2024–10, 2024.

[77] Guillaume Huguet, Daniel Sumner Magruder, Alexander Tong, Oluwadamilola Fasina, Manik Kuchroo, Guy Wolf, and Smita Krishnaswamy. Manifold interpolating optimal-transport flows for trajectory inference. *Advances in neural information processing systems*, 35:29705–29718, 2022.

[78] Geert-Jan Huizing, Gabriel Peyré, and Laura Cantini. Optimal transport improves cell–cell similarity inference in single-cell omics data. 38(8):2169–2177.

[79] Yongxia Huo, Shiwu Li, Jiewei Liu, Xiaoyan Li, and Xiong-Jian Luo. Functional genomics reveal gene regulatory mechanisms underlying schizophrenia risk. *Nature communications*, 10(1):670, 2019.

[80] Yasser Iturria-Medina, Quadri Adewale, Ahmed F Khan, Simon Ducharme, Pedro Rosa-Neto, Kieran O'Donnell, Vladislav A Petyuk, Serge Gauthier, Philip L De Jager, John Breitner, et al. Unified epigenomic, transcriptomic, proteomic, and metabolomic taxonomy of alzheimer's disease progression and heterogeneity. *Science advances*, 8(46):eabo6764, 2022.

[81] Qi Jiang and Lin Wan. A physics-informed neural sde network for learning cellular dynamics from time-series scrna-seq data. *Bioinformatics*, 40(Supplement_2):ii120–ii127, 2024.

[82] Suoqin Jin, Christian F Guerrero-Juarez, Lihua Zhang, Ivan Chang, Raul Ramos, Chen-Hsiang Kuan, Peggy Myung, Maksim V Plikus, and Qing Nie. Inference and analysis of cell-cell communication using cellchat. *Nature communications*, 12(1):1088, 2021.

[83] L. Kantorovitch. On the translocation of masses.

[84] Arielle S Keller, Adam R Pines, Sheila Shanmugan, Valerie J Sydnor, Zaixu Cui, Maxwell A Bertolero, Ran Barzilay, Aaron F Alexander-Bloch, Nora Byington, Andrew Chen, et al. Personalized functional brain network topography is associated with individual differences in youth cognition. *Nature communications*, 14(1):8411, 2023.

[85] J L Kennedy, L A Farrer, N C Andreasen, R Mayeux, and P St George-Hyslop. The genetics of Adult-Onset neuropsychiatric disease: Complexities and conundra? *Science*, 302:822–826, 2003.

[86] Beomsu Kim, Gihyun Kwon, Kwanyoung Kim, and Jong Chul Ye. Unpaired image-to-image translation via neural schr\" odinger bridge. *arXiv preprint arXiv:2305.15086*, 2023.

[87] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[88] P Kodam, R Sai Swaroop, S S Pradhan, V Sivaramakrishnan, and R Vadrevu. Integrated multi-omics analysis of alzheimer's disease shows molecular signatures associated with disease progression and potential therapeutic targets. *Sci. Rep*, 13, 2023.

[89] Solomon Kullback. Probability densities with given marginals. *The Annals of Mathematical Statistics*, 39(4):1236–1243, 1968.

[90] Jean-Charles Lambert, Simon Heath, Gael Even, Dominique Campion, Kristel Sleegers, Mikko Hiltunen, Onofre Combarros, Diana Zelenika, Maria J Bullido, Béatrice Tavernier, et al. Genome-wide association study identifies variants at clu and cr1 associated with alzheimer's disease. *Nature genetics*, 41(10):1094–1099, 2009.

[91] Allan M Landes, Susan D Sperry, and Milton E Strauss. Prevalence of apathy, dysphoria, and depression in relation to dementia severity in alzheimer's disease. *The Journal of neuropsychiatry and clinical neurosciences*, 17(3):342–349, 2005.

[92] Donghoon Lee, Mikaela Koutrouli, Nicolas Y Masse, Gabriel E Hoffman, Seon Kinrot, Xinyi Wang, Prashant, Milos Pjanic, Tereza Clarence, Fotios Tsetsos, Deepika Mathur, David Burstein, Karen Therrien, Aram Hong, Clara Casey, Zhiping Shao, Marcela Alvia, Stathis Argyriou, Jennifer Monteiro Fortes, Pavel Katsel, Pavan K Auluck, Lisa L Barnes, Stefano Marenco, David A Bennett, PsychAD Consortium, Lars Juhl Jensen, Kiran Girdhar, Georgios Voloudakis, Vahram Haroutunian, Jaroslav Bendl, John F Fullard, and Panos Roussos. Single-cell atlas of transcriptomic vulnerability across multiple neurodegenerative and neuropsychiatric diseases. November 2024.

[93] Mark DM Leiserson, Fabio Vandin, Hsin-Ta Wu, Jason R Dobson, Jonathan V Eldridge, Jacob L Thomas, Alexandra Papoutsaki, Younhun Kim, Beifang Niu, Michael McLellan, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature genetics*, 47(2):106–114, 2015.

[94] Christian L'eonard. A survey of the schrödinger problem and some of its connections with optimal transport. *arXiv: Probability*, 2013.

[95] Esten H Leonardsen, Karin Persson, Edvard Grødem, Nicola Dinsdale, Till Schellhorn, James M Roe, Didac Vidal-Piñeiro, Øystein Sørensen, Tobias Kaufmann, Eric Westman, et al. Constructing personalized characterizations of structural brain aberrations in patients with dementia using explainable artificial intelligence. *NPJ Digital Medicine*, 7(1):110, 2024.

[96] Ganna Leonenko, Emily Baker, Joshua Stevenson-Hoare, Annerieke Sierksma, Mark Fiers, Julie Williams, Bart De Strooper, and Valentina Escott-Price. Identifying individuals with high risk of alzheimer's disease using polygenic risk scores. *Nature communications*, 12(1):4506, 2021.

[97] Hong Lian, Alexandra Litvinchuk, Angie C-A Chiang, Nadia Aithmitti, Joanna L Jankowsky, and Hui Zheng. Astrocyte-microglia cross talk through complement activation modulates amyloid pathology in mouse models of alzheimer's disease. *Journal of Neuroscience*, 36(2):577–589, 2016.

[98] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell systems*, 1(6):417–425, 2015.

[99] Agnes Lindbo, Maria Gustafsson, Ulf Isaksson, Per-Olof Sandman, and Hugo Lövheim. Dysphoric symptoms in relation to other behavioral and psychological symptoms of dementia, among elderly in nursing homes. *BMC geriatrics*, 17:1–8, 2017.

[100] Jie Liu, Yuanhao Huang, Ritambhara Singh, Jean-Philippe Vert, and William Stafford Noble. Jointly embedding multiple single-cell omics measurements.

[101] Longqi Liu, Chuanyu Liu, Andrés Quintero, Liang Wu, Yue Yuan, Mingyue Wang, Mengnan Cheng, Lizhi Leng, Liqin Xu, Guoyi Dong, Rui Li, Yang Liu, Xiaoyu Wei, Jiangshan Xu, Xiaowei Chen, Haorong Lu, Dongsheng Chen, Quanlei Wang, Qing Zhou, Xinxin Lin, Guibo Li, Shiping Liu, Qi Wang, Hongru Wang, J. Lynn Fink, Zhengliang Gao, Xin Liu, Yong Hou, Shida Zhu, Huanming Yang, Yunming Ye, Ge Lin, Fang Chen, Carl Herrmann, Roland Eils, Zhouchun Shang, and Xun Xu. Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. 10(1):470.

[102] Xiaoping Liu, Yuetong Wang, Hongbin Ji, Kazuyuki Aihara, and Luonan Chen. Personalized characterization of diseases using sample-specific networks. *Nucleic acids research*, 44(22):e164–e164, 2016.

[103] Nick Z. Lu, Suzanne E. Wardell, Kerry L. Burnstein, Donald Defranco, Peter J. Fuller, Vincent Giguere, Richard B. Hochberg, Lorraine McKay, Jack-Michel Renoir, Nancy L. Weigel, Elizabeth M. Wilson, Donald P. McDonnell, and John A. Cidlowski. International union of pharmacology. LXV. the pharmacology and classification of the nuclear receptor superfamily: Glucocorticoid, mineralocorticoid, progesterone, and androgen receptors. 58(4):782–797.

[104] Shan Lu and Sündüz Keleş. Debiased personalized gene coexpression networks for population-scale scrna-seq data. *Genome Research*, 33(6):932–947, 2023.

[105] Anjun Ma, Xiaoying Wang, Jingxian Li, Cankun Wang, Tong Xiao, Yuntao Liu, Hao Cheng, Juexin Wang, Yang Li, Yuzhou Chang, et al. Single-cell biological network inference using a heterogeneous graph transformer. *Nature Communications*, 14(1):964, 2023.

[106] Sophie E Mastenbroek, Jacob W Vogel, Lyduine E Collij, Geidy E Serrano, Cécilia Tremblay, Alexandra L Young, Richard A Arce, Holly A Shill, Erika D Driver-Dunckley, Shyamal H Mehta, et al. Disease progression modelling reveals heterogeneity in trajectories of lewy-type α-synuclein pathology. *Nature communications*, 15(1):5133, 2024.

[107] Hansruedi Mathys, Carles A Boix, Leyla Anne Akay, Ziting Xia, Jose Davila-Velderrain, Ayesha P Ng, Xueqiao Jiang, Ghada Abdelhady, Kyriaki Galani, Julio Mantero, et al. Single-cell multiregion dissection of alzheimer's disease. *Nature*, 632(8026):858–868, 2024.

[108] Hansruedi Mathys, Jose Davila-Velderrain, Zhuyu Peng, Fan Gao, Shahin Mohammadi, Jennie Z Young, Madhvi Menon, Liang He, Fatema Abdurrob, Xueqiao Jiang, et al. Single-cell transcriptomic analysis of alzheimer's disease. *Nature*, 570(7761):332–337, 2019.

[109] Hansruedi Mathys, Zhuyu Peng, Carles A Boix, Matheus B Victor, Noelle Leary, Sudhagar Babu, Ghada Abdelhady, Xueqiao Jiang, Ayesha P Ng, Kimia Ghafari, et al. Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to alzheimer's disease pathology. *Cell*, 186(20):4365–4385, 2023.

[110] María Luisa Menéndez, Julio Angel Pardo, Leandro Pardo, and María del C Pardo. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318, 1997.

[111] Lucia Migliore and Fabio Coppedè. Gene–environment interactions in alzheimer disease: the emerging role of epigenetics. *Nature Reviews Neurology*, 18(11):643–660, 2022.

[112] Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. 11(4):417–487.

[113] Ryan A Neff, Minghui Wang, Sezen Vatansever, Lei Guo, Chen Ming, Qian Wang, Erming Wang, Emrin Horgusluoglu-Moloch, Won-min Song, Aiqun Li, et al. Molecular subtyping of alzheimer's disease using rna sequencing data reveals novel mechanisms and targets. *Science advances*, 7(2):eabb5398, 2021.

[114] Nam D. Nguyen, Ian K. Blaby, and Daifeng Wang. ManiNetCluster: a novel manifold learning approach to reveal the functional links between gene networks. 20:1003.

[115] Sid E O'Bryant, Stephen C Waring, C Munro Cullum, James Hall, Laura Lacritz, Paul J Massman, Philip J Lupo, Joan S Reisch, Rachelle Doody, Texas Alzheimer's Research Consortium, et al. Staging dementia using clinical dementia rating scale sum of boxes scores: a texas alzheimer's research consortium study. *Archives of neurology*, 65(8):1091–1095, 2008.

[116] Jorge J Palop and Lennart Mucke. Synaptic depression and aberrant excitatory network activity in alzheimer's disease: two faces of the same coin? *Neuromolecular medicine*, 12:48–55, 2010.

[117] Matteo Pariset, Ya-Ping Hsieh, Charlotte Bunne, Andreas Krause, and Valentin De Bortoli. Unbalanced diffusion schrödinger bridge. *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, abs/2306.09099, 2023.

[118] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[119] Ronald Carl Petersen, Paul S Aisen, Laurel A Beckett, Michael C Donohue, Anthony Collins Gamst, Danielle J Harvey, CR Jack Jr, William J Jagust, Leslie M Shaw, Arthur W Toga, et al. Alzheimer's disease neuroimaging initiative (adni) clinical characterization. *Neurology*, 74(3):201–209, 2010.

[120] Gabriel Peyré and Marco Cuturi. Computational optimal transport.

[121] Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, Javier García-García, Ferran Sanz, and Laura I Furlong. Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research*, page gkw943, 2016.

[122] Milton Pividori, Sumei Lu, Binglan Li, Chun Su, Matthew E Johnson, Wei-Qi Wei, Qiping Feng, Bahram Namjou, Krzysztof Kiryluk, Iftikhar J Kullo, et al. Projecting genetic associations through gene expression patterns highlights disease etiology and drug mechanisms. *Nature communications*, 14(1):5562, 2023.

[123] Xiaojie Qiu, Yan Zhang, Jorge D. Martin-Rufino, Chen Weng, Shayan Hosseinzadeh, Dian Yang, Angela N. Pogson, Marco Y. Hein, Kyung Hoi (Joseph) Min, Li Wang, Emanuelle I. Grody, Matthew J. Shurtleff, Ruoshi Yuan, Song Xu, Yian Ma, Joseph M. Replogle, Eric S. Lander, Spyros Darmanis, Ivet Bahar, Vijay G. Sankaran, Jianhua Xing, and Jonathan S. Weissman. Mapping transcriptomic vector fields of single cells. *Cell*, 185(4):690–711.e45, 2022.

[124] Timothy E. Reddy, Jason Gertz, Gregory E. Crawford, Michael J. Garabedian, and Richard M. Myers. The hypersensitive glucocorticoid response specifically regulates period 1 and expression of circadian genes. 32(18):3756–3767.

[125] Timothy E. Reddy, Florencia Pauli, Rebekka O. Sprouse, Norma F. Neff, Kimberly M. Newberry, Michael J. Garabedian, and Richard M. Myers. Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. 19(12):2163–2171.

[126] MORRIS W RIRSCH and Stephen Smale. *Differential equations, dynamical systems, and linear algebra*. ACADEMIC PRESS. INC., 1974.

[127] JJ Rodríguez-Arellano, Vladimir Parpura, Robert Zorec, and Alexei Verkhratsky. Astrocytes in physiological aging and alzheimer's disease. *Neuroscience*, 323:170–182, 2016.

[128] Adria Ruiz, Oriol Martinez, Xavier Binefa, and Jakob Verbeek. Learning disentangled representations with reference-based variational autoencoders. Publisher: arXiv Version Number: 1.

[129] Ludger Rüschendorf. Convergence of the iterative proportional fitting procedure. *The Annals of Statistics*, pages 1160–1174, 1995.

[130] Alexandre Gramfort Mokhtar Z. Alaya Aurélie Boisbunon Stanislas Chambon Laetitia Chapel Adrien Corenflos Kilian Fatras Nemo Fournier Léo Gautheron Nathalie T.H. Gayraud Hicham Janati Alain Rakotomamonjy Ievgen Redko Antoine Rolet Antony Schutz Vivien Seguy Danica J. Sutherland Romain Tavenard Alexander Tong Titouan Vayer Rémi Flamary, Nicolas Courty. POT python optimal transport library.

[131] Lazaro M Sanchez-Rodriguez, Gleb Bezgin, Felix Carbonell, Joseph Therriault, Jaime Fernandez-Arias, Stijn Servaes, Nesrine Rahmouni, Cécile Tissot, Jenna Stevenson, Thomas K Karikari, et al. Personalized whole-brain neural mass models reveal combined aβ and tau hyperexcitable influences in alzheimer's disease. *Communications Biology*, 7(1):528, 2024.

[132] Andrew Octavian Sasmita, Constanze Depp, Taisiia Nazarenko, Ting Sun, Sophie B Siems, Erinne Cherisse Ong, Yakum B Nkeh, Carolin Böhler, Xuan Yu, Bastian Bues, et al. Oligodendrocytes produce amyloid-β and contribute to plaque formation alongside neurons in alzheimer's disease model mice. *Nature neuroscience*, 27(9):1668–1674, 2024.

[133] Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

[134] Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, Lia Lee, Jenny Chen, Justin Brumbaugh, Philippe Rigollet, Konrad Hochedlinger, Rudolf Jaenisch, Aviv Regev, and Eric S. Lander. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. 176(4):928–943.e22.

[135] Alberto Serrano-Pozo, Huan Li, Zhaozhi Li, Clara Muñoz-Castro, Methasit Jaisa-Aad, Molly A Healey, Lindsay A Welikovitch, Rojashree Jayakumar, Annie G Bryant, Ayush Noori, et al. Astrocyte transcriptomic changes along the spatiotemporal progression of alzheimer's disease. *Nature neuroscience*, pages 1–17, 2024.

[136] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[137] You-Hyang Song, Jiwon Yoon, and Seung-Hee Lee. The role of neuropeptide somatostatin in the brain and its application in treating neurological disorders. *Experimental & Molecular Medicine*, 53(3):328–338, 2021.

[138] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. 14(9):865–868.

[139] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *cell*, 177(7):1888–1902, 2019.

[140] Shengbao Suo, Qian Zhu, Assieh Saadatpour, Lijiang Fei, Guoji Guo, and Guo-Cheng Yuan. Revealing the critical regulators of cell identity in the mouse cell atlas. *Cell reports*, 25(6):1436–1445, 2018.

[141] Lukasz Mateusz Szewczyk, Marcin Andrzej Lipiec, Ewa Liszewska, Ksenia Meyza, Joanna Urban-Ciecko, Ludwika Kondrakiewicz, Anna Goncerzewicz, Kamil Rafalko, Tomasz Grzegorz Krawczyk, Karolina Bogaj, et al. Astrocytic β-catenin signaling via tcf7l2 regulates synapse development and social behavior. *Molecular Psychiatry*, 29(1):57–73, 2024.

[142] Thibault Séjourné, François-Xavier Vialard, and Gabriel Peyré. The unbalanced gromov wasserstein distance: Conic formulation and relaxation.

[143] Alice Taubes, Phil Nova, Kelly A Zalocusky, Idit Kosti, Mesude Bicak, Misha Y Zilberter, Yanxia Hao, Seo Yeon Yoon, Tomiko Oskotsky, Silvia Pineda, et al. Experimental and real-world evidence supporting the computational repurposing of bumetanide for apoe4-related alzheimer's disease. *Nature Aging*, 1(10):932–947, 2021.

[144] Dietmar R Thal, Udo Rub, Mario Orantes, and Heiko Braak. Phases of aβ-deposition in the human brain and its relevance for the development of ad. *Neurology*, 58(12):1791–1800, 2002.

[145] Alexander Tong, Jessie Huang, Guy Wolf, David Van Dijk, and Smita Krishnaswamy. Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. In *International conference on machine learning*, pages 9526–9536. PMLR, 2020.

[146] Leonardo Tozzi, Xue Zhang, Adam Pines, Alisa M Olmsted, Emily S Zhai, Esther T Anene, Megan Chesnut, Bailey Holt-Gosselin, Sarah Chang, Patrick C Stetz, et al. Personalized brain circuit scores identify clinically distinct biotypes in depression and anxiety. *Nature medicine*, 30(7):2076–2087, 2024.

[147] Duc Tran, Hung Nguyen, Bang Tran, Carlo La Vecchia, Hung N Luu, and Tin Nguyen. Fast and precise single-cell data analysis using a hierarchical autoencoder. *Nature communications*, 12(1):1029, 2021.

[148] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J. Lennon, Kenneth J. Livak, Tarjei Sigurd Mikkelsen, and John L. Rinn. Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. *Nature biotechnology*, 32:381 – 386, 2014.

[149] Alexandro E. Trevino, Fabian Müller, Jimena Andersen, Laksshman Sundaram, Arwa Kathiria, Anna Shcherbina, Kyle Farh, Howard Y. Chang, Anca M. Pașca, Anshul Kundaje, Sergiu P. Pașca, and William J. Greenleaf. Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. 184(19):5053–5069.e23.

[150] Monique GP Van Der Wijst, Dylan H de Vries, Harm Brugge, Harm-Jan Westra, and Lude Franke. An integrative approach for building personalized gene regulatory networks for precision medicine. *Genome medicine*, 10:1–15, 2018.

[151] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[152] Adrian Veres, Aubrey L Faust, Henry L Bushnell, Elise N Engquist, Jennifer Hyoje-Ryu Kenty, George Harb, Yeh-Chuin Poh, Elad Sintov, Mads Gürtler, Felicia W Pagliuca, et al. Charting cellular identity during human in vitro β-cell differentiation. *Nature*, 569(7756):368–373, 2019.

[153] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

[154] Lei Wan, Ping Zhong, Pei Li, Yong Ren, Wei Wang, Mingjun Yu, Henry Y Feng, and Zhen Yan. Crispr-based epigenetic editing of gad1 improves synaptic inhibition and cognitive behavior in a tauopathy mouse model. *Neurobiology of Disease*, page 106826, 2025.

[155] Daifeng Wang. Comprehensive functional genomic resource and integrative model for the adult brain. In *NEUROPSYCHOPHARMACOLOGY*, volume 43, pages S16–S16. NATURE PUBLISHING GROUP MACMILLAN BUILDING, 4 CRINAN ST, LONDON N1 9XW, ENGLAND, 2018.

[156] Gefei Wang, Yuling Jiao, Qian Xu, Yang Wang, and Can Yang. Deep generative learning via schrödinger bridge. In *International conference on machine learning*, pages 10794–10804. PMLR, 2021.

[157] Juexin Wang, Anjun Ma, Yuzhou Chang, Jianting Gong, Yuexu Jiang, Ren Qi, Cankun Wang, Hongjun Fu, Qin Ma, and Dong Xu. scgnn is a novel graph neural network framework for single-cell rna-seq analyses. *Nature communications*, 12(1):1882, 2021.

[158] Qi Wang, Jerry Antone, Eric Alsop, Rebecca Reiman, Cory Funk, Jaroslav Bendl, Joel T Dudley, Winnie S Liang, Timothy L Karr, Panos Roussos, et al. Single cell transcriptomes and multiscale networks from persons with and without alzheimer's disease. *Nature communications*, 15(1):5815, 2024.

[159] Caleb Weinreb, Alejo Rodriguez-Fraticelli, Fernando D Camargo, and Allon M Klein. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science*, 367(6479):eaaw3381, 2020.

[160] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.

[161] F. Alexander Wolf, Fiona Hamey, Mireya Plass, Jordi Solana, Joakim S. Dahlin, Berthold Göttgens, Nikolaus Rajewsky, Lukas Simon, and Fabian J. Theis. Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *bioRxiv*, 2018.

[162] Shuyang Yao, Arvid Harder, Fahimeh Darki, Yu-Wei Chang, Ang Li, Kasra Nikouei, Giovanni Volpe, Johan N Lundström, Jian Zeng, Naomi R Wray, et al. Connecting genomic results for psychiatric disorders to human brain cell types and regions reveals convergence with functional connectivity. *Nature Communications*, 16(1):395, 2025.

[163] Grace Hui Ting Yeo, Sachit D Saksena, and David K Gifford. Generative modeling of single-cell time series with prescient enables prediction of cell trajectories with interventions. *Nature communications*, 12(1):3222, 2021.

[164] A L Young. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference. *Nat. Commun*, 9, 2018.

[165] Qinggan Zeng, Rongyong Man, Yifeng Luo, Ling Zeng, Yushi Zhong, Bingxun Lu, and Xiaofeng Wang. Irf-8 is involved in amyloid-β 1–40 (aβ 1–40)-induced microglial activation: a new implication in alzheimer's disease. *Journal of Molecular Neuroscience*, 63:159–164, 2017.

[166] Jiaqi Zhang, Erica Larschan, Jeremy Bigness, and Ritambhara Singh. scnode: generative model for temporal single cell transcriptomic data prediction. *Bioinformatics*, 40(Supplement_2):ii146–ii154, 2024.

[167] Ran Zhang, Laetitia Meng-Papaxanthos, Jean-Philippe Vert, and William Stafford Noble. Semi-supervised single-cell cross-modality translation using polarbear.

[168] Yingyao Zhou, Bin Zhou, Lars Pache, Max Chang, Alireza Hadj Khodabakhshi, Olga Tanaseichuk, Christopher Benner, and Sumit K. Chanda. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. 10(1):1523.

[169] Yingyue Zhou, Wilbur M Song, Prabhakar S Andhey, Amanda Swain, Tyler Levy, Kelly R Miller, Pietro L Poliani, Manuela Cominelli, Shikha Grover, Susan Gilfillan, et al. Human and mouse single-nucleus transcriptomics reveal trem2-dependent and trem2-independent cellular responses in alzheimer's disease. *Nature medicine*, 26(1):131–142, 2020.

[170] M N Ziats, L P Grosvenor, and O M Rennert. Functional genomics of human brain development and implications for autism spectrum disorders. *Transl. Psychiatry*, 5:e665–e665, 2015.