Using hierarchical variational autoencoders to incorporate conditional independent priors for paired single-cell multi-omics data integration

Ping-Han Hsieh
Centre for Molecular Medicine Norway
Department of Informatics
University of Oslo
pinghanh@ncmm.uio.no

Ru-Xiu Hsiao
Genome and Systems Biology Degree Program
National Taiwan University
r09b48004@ntu.edu.tw

Tatiana Belova
Centre for Molecular Medicine Norway
University of Oslo
tatiana.belova@ncmm.uio.no

Katalin Ferenc
Centre for Molecular Medicine Norway
University of Oslo
k.t.ferenc@ncmm.uio.no

Anthony Mathelier
Centre for Molecular Medicine Norway
University of Oslo
anthony.mathelier@ncmm.uio.no

Rebekka Burkholz
CISPA Helmholtz Center for Information Security
burkholz@cispa.de

Chien-Yu Chen
Department of Biomechatronics Engineering
Genome and Systems Biology Degree Program
National Taiwan University
chienyuchen@ntu.edu.tw

Geir Kjetil Sandve
Department of Informatics
University of Oslo
geirksa@ifi.uio.no

Marieke Kuijjer
Centre for Molecular Medicine Norway
University of Oslo
Department of Pathology
Leiden University Medical Center
marieke.kuijjer@ncmm.uio.no

Abstract

Recently, paired single-cell sequencing technologies have allowed the measurement of multiple modalities of molecular data simultaneously, at single-cell resolution. Along with the advances in these technologies, many methods based on variational autoencoder have been developed aiming at integrating paired single-cell multi-omics data. However, how to incorporate prior biological understanding of data properties into such models remains an open question in the field. Here, we propose a novel probabilistic learning framework that explicitly incorporates conditional independence relationships between multi-modal data as a directed acyclic graph using a generalized hierarchical variational autoencoder. Applying our approach to single-cell ATAC and RNA-seq data, we find that our method can identify cell clusters with distinct expression profiles that are not driven by chromatin state. We anticipate that our proposed framework can help construct flexible graphical models

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

that reflect biological hypotheses with ease and unravel the interactions between different biological data types, such as different modalities of paired single-cell multi-omics data. The implementation of the proposed framework can be found in the repository `https://github.com/kuijjerlab/CAVACHON`.

1 Introduction

In the past decade, the emergence of high throughput single-cell single-omics technologies has enabled the profiling of snapshots of gene and protein expression [1, 2, 3], chromatin accessibility [4, 5, 6] and transcription factor binding [7] occurring in each individual cell. Single-cell datasets generated with these technologies have broaden our understanding of, for example, stochasticity of gene expression [8], expression dynamics during cell differentiation [9], and differences in chromatin states between cells from healthy and diseased tissues [10, 11]. However, as regulatory elements work cooperatively to mediate biological processes in cells, single-omics technologies may fail to fully capture the complex molecular mechanisms involving multiple dimensions or modalities of regulatory elements [12].

Thanks to recent advances in single-cell multi-omics sequencing technology such as CITE-Seq [13], scNMT-Seq [14], SNARE-Seq [15], SHARE-Seq [16] and 10X Multiome [17], it is now possible to capture multiple modalities of molecular data simultaneously at single-cell resolution. However, integrating and deriving biological insights from multiple modalities is challenging, as each modality requires distinct data processing, normalization, modeling and interpretation [12]. To address this, various data analysis tools and computational methods have been proposed. For instance, Seurat [18], Harmony [19], Scanpy [20] and MUON [21] all provide unified frameworks for quality control, data exploration and downstream analysis of single-cell data. Other methods such as scMVAE [22], Cobolt [23], scMM [24] and TotalVI [25] integrate multiple modalities of single-cell data by applying variational autoencoders to embed the observed data into a joint low-dimensional latent space. However, using a joint latent space renders it difficult to interpret the model and unravel the interactions between different modalities. Moreover, most of these published methodologies aim at direct multi-omics data integration. How to incorporate prior biological understanding of the properties of data into the existing model remains an open question in the field.

Recently, GLUE [26] was proposed to incorporate a prior knowledge graph between biological features to link the generative process between different modalities. However, this approach required constructing specific interactions between regulatory elements in advance, which might render it difficult to apply when the understanding of the detailed underlying regulatory network is lacking. Complementary, we propose a novel probabilistic learning framework that incorporates the conditional independence relationships on the modality-level (e.g. state of chromatin accessibility may influence gene expression in general, as opposed to specific chromatin region may influence specific gene expression) using a generalized variational ladder autoencoder [27]. Specifically, our proposed framework takes as input the observed data sets and a directed acyclic graph (optional) representing the relationships between modalities. Then, the model is sequentially trained based on the topological order in the provided graph to optimize the evidence lower bound of the data likelihood of each modality. Applying to the SNARE-Seq data from the cerebral cortex of an adult mouse, we showed that our model is capable of isolating explained variability of gene expression from chromatin accessibility signals. We anticipate that our proposed framework will help users to construct flexible graphical models that reflect biological hypotheses with ease and unravel the interactions between different biological data types, such as different modalities of paired single-cell multi-omics data.

2 Methods

2.1 Problem Definition

Consider M multiple modalities of data with the same anchor, and a directed acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_m \mid m = 1, 2...M\}$ is the set of vertices representing the data modalities, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of unweighted directed edges specifying our prior knowledge on the conditional independence relationships between different modalities. Conditional independence relationships between different modalities are provided as follows: If there exists a directed edge from modality A to modality B, then modality A and modality B are conditionally independent given the latent

representation of modality A (see Figure 1). Our proposed method aims to learn the generative process of the observed data set given the dependency between the latent distributions across different modalities by creating and training a hierarchical variational autoencoder explicitly based on $\mathcal{G}$.

For paired single-cell multi-omics sequencing data, each modality represents a single molecular assay anchored by the barcodes of cells. The graphical prior knowledge provided to our method could, for example, be given by the states of chromatin accessibility that may facilitate regulation of gene expression [28] (Figure 1b). Note that although we mainly present our proposed method in the context of paired single-cell multi-omics data, it can also be applied to single-omics data derived from single-cells or bulk tissues, e.g. time series data analysis (Figure 1c). For other potential use cases, please refer to Figure 1.

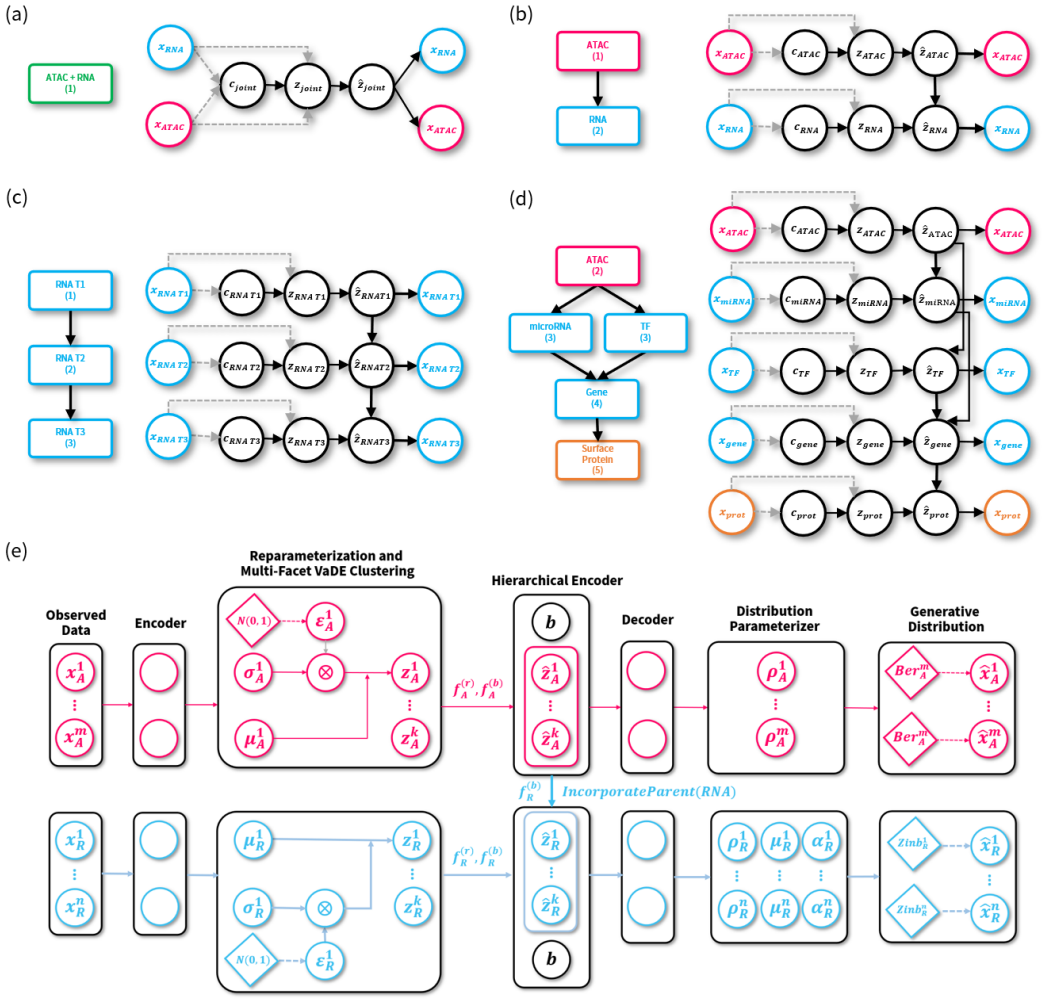

Figure 1: Examples of relational graphs and graphical models. (a)-(d) Schematic diagrams of input graphs representing conditional independent relationships (left) between modalities of data and the graphical models created by our method (right). Note that the batch information is omitted in (a)-(d). The numbers under the names of the rectangle boxes are the topological orders used for sequential training. The nodes in the graphical model are colored by the type of molecular assays. The dashed arrows denote the recognition model for posterior approximation while the solid arrows denote the generative process. The figure shows examples with (a) joint learning, (b) states of chromatin accessibility that may influence the expression of genes, (c) gene expression of later time points in a time series experiment are dependent on earlier time points, (d) the states of chromatin accessibility may control how transcription factors and microRNAs regulate gene expression, and eventually affect the abundance of surface proteins. (e) The architecture of the created hierarchical variational autoencoder for the graphical model of (b).

2.2 Probabilistic Generative Model

Inspired by a recently published method designed for multi-facet clustering in computer vision [29], we applied a generalized version of a variational ladder autoencoder to incorporate prior knowledge of the conditional independence relationships between biological data into a probabilistic learning framework. This approach allows multiple cluster assignments that conditionally separate the abstract features of each modality of the data, based on a provided relational graph. This is particularly helpful for identifying new clusters of cell types. Our proposed model approximates the generative process of the observed multimodal data explicitly based on the given prior conditional independence relationships between modalities (Equation 1).

$$\begin{aligned} c_m &\sim Categorical(\pi_m) \\ z_m &\sim IndependentNormal(\mu(c_m), \Sigma(c_m)) \\ \tilde{z}_m &= f_m^{(r)}(z_m; \theta_m^{(r)}) \\ \hat{z}_m &= IncorporateParents(m) \\ \rho_m &= f_m^{(d)}([\hat{z}_m; b]; \theta_m^{(d)}) \\ x_m &\sim Dist(\rho_m), \end{aligned} \tag{1}$$

where c_m is the cluster assignment for the mixture of independent Gaussian distributions with the corresponding mean $\mu(c_m)$ and standard deviation $\Sigma(c_m)$ and b is the batch information that needs to be corrected for the anchors. We correct for batch effects by conditioning the decoder on b [30, 31]. z_m is the latent representation of modality m. $IncorporateParents$ is a function that incorporates the latent representations of the parents of the m-th modality $\mathcal{P}(m) = \{p|(p, m) \in \mathcal{E}\}$ into the posterior approximation considering the prior conditional independence relationships specified in $\mathcal{G}$. Here, we concatenate the latent representations of the parent modalities and condition the posterior distribution on them as follows:

$$IncorporateParents(m) = f_m^{(b)}([\hat{z}_{\mathcal{P}}; \tilde{z}_m]; \theta_m^{(b)}), \tag{2}$$

where $f_m^{(b)}$ and $f_m^{(r)}$ are linear transformations with parameters $\theta_m^{(b)}$ and $\theta_m^{(r)}$. $f_m^{(d)}$ is the decoder neural network used to parameterize the specified data distribution. ρ_m is the set of parameters of user-defined data distributions $Dist$ used to compute the data likelihood. A more detailed description of the framework and neural architecture of $f_m^{(b)}$, $f_m^{(r)}$, $f_m^{(e)}$ and $f_m^{(d)}$ is shown in Appendix Table S1. The generative process of m-th modality can be structured as $p_\theta^{(m)}(x_m) = p_\theta^{(m)}(x_m|z_m, z_{\mathcal{A}(m)}, b)p_\theta^{(m)}(z_m|c_m)p_\theta^{(m)}(c_m)$, where $\mathcal{A}(m)$ is the set of ancestors of modality m (Appendix A).

2.3 Training Strategy

Previous studies have shown that *progressive training* improves the ability of the model to learn disentangled representations across layers in hierarchical variational autoencoders [29, 32]. We found that this is particularly helpful to prevent the offspring modality from learning redundant representations that have been encoded in the latent representations of their ancestors.

We applied *sequential training* [33] based on the order of modalities (e.g. the numbers shown under the names of the components in Figure 1) after the topological sort of the provided graph G. Specifically, when training the model of the m-th modality with its ancestors $\mathcal{A}(m)$ and their offspring $\mathcal{O}(m)$, we fixed the trainable weights in $f^{(e)}_{\mathcal{A}(m)\cup\mathcal{O}(m)}$, $f^{(d)}_{\mathcal{A}(m)\cup\mathcal{O}(m)}$, $f^{(r)}_{\mathcal{A}(m)\cup\mathcal{O}(m)}$ and $f^{(b)}_{\mathcal{A}(m)\cup\mathcal{O}(m)}$. This way, the posteriors and the generative processes for the ancestor and offspring modalities remain unchanged (Figure 2). This complements progressive training and prevents ancestor modalities from learning representations that are only relevant to the offspring. As shown in previous research [33], sequential training also improves the training stability and reduces the memory usage.

For the analysis presented in this study, we trained the model for 1,500 epochs in each stage with early stopping using the Adam optimizer [34] with a learning rate of 1e-4. The model is trained by maximizing the evidence lower bound (ELBO) of the conditional data likelihood. For the set consisting of all modalities $\mathcal{V}$, the ELBO is computed as follows:

$$\log p(x_{\mathcal{V}}|b) \geq \mathbb{E}_{q_\phi(z_{\mathcal{V}},c_{\mathcal{V}}|x_{\mathcal{V}},b)}[\log(\frac{p_\theta(x_{\mathcal{V}},z_{\mathcal{V}},c_{\mathcal{V}}|b)}{q_\phi(z_{\mathcal{V}},c_{\mathcal{V}}|x_{\mathcal{V}},b)})] = \mathcal{L}(x_{\mathcal{V}},b;\theta,\phi), \tag{3}$$

where $q_\phi(z_{\mathcal{V}}, c_{\mathcal{V}}|x_{\mathcal{V}}, b)$ is parameterized by encoder neural networks $f_{\mathcal{V}}^{(e)}$. For multi-facet online clustering, we applied the single-facet VaDE trick [29, 35] sequentially to identify the cluster assignments of each data point. For a more detailed derivation, see Appendices A and B. For the analysis shown in this study, the dimensionalities of the latent distribution is set to $\dim(z) = 20$. To ensure the model is identifiable, we set the number of components in the mixture of independent Gaussian priors to $K = 2 \times \dim(z) + 1 = 41$ [36, 37].

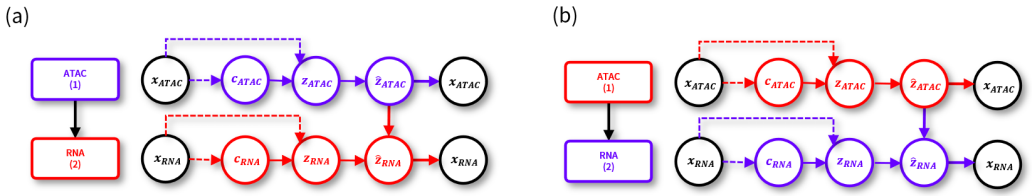

Figure 2: Sequential training strategy. (a) the first stage and (b) the second stage of sequential training. On the left panel, the purple box suggests the component to optimize, while the red box is the component with frozen trainable weights in each stage. On the right panel, the purple arrows suggest regular forward and backpropagation, while the red arrows suggest the gradients of the ELBO with respect to the trainable weights in these functions that are frozen during backpropagation.

2.4 Datasets

For the analysis presented in this study, we used SNARE-Seq data from the cerebral cortex of an adult mouse (accession number GSE126074) [15]. These data include modalities on both chromatin state and gene expression in the same cell. We binarized the chromatin accessibility data and used the Bernoulli distribution to model the data distribution. For the gene expression data, we used the zero-inflated negative binomial distribution, similar to what was done in previous studies [30, 31]. We annotated cell types using marker genes from the single-cell transcriptomic reference data set of the mouse visual cortex from the Allen Brain Atlas [38]. We identified these genes using the *FindTransferAnchors* function in Bioconductor package Seurat (version 4.1.0) [18]. More details on the parameters we used during preprocessing can be found in Appendix D.

3 Results

3.1 Conditional Latent Representation

Here, we present some use cases of our framework. As an example, we hypothesize that chromatin accessibility states influence the regulation of gene expression. Therefore, the corresponding hierarchical variational autoencoder is explicitly created based on our hypothesis between modalities (Figure 1b, e). We applied the created model to SNARE-Seq data obtained from the cerebral cortex of an adult mouse, which includes paired chromatin accessibility states and gene expression information in each single cell (see Method 2.4 for more details). We showed that our method is capable of isolating the source of variability based on the prior knowledge graph between modalities, and that the latent representation of gene expression would only capture the residual of the explained variability considering the states of chromatin accessibility. (Figure 3c, 3d). In particular, our model identifies distinct clusters of astrocytes (Astro), nucleus pulposus cells (NP) and Pvalb neurons (Pvalb) in the conditional latent representations of gene expression data that consider the observed chromatin accessibility (Figure 3b). This implies that the distinction of these cells from other cell types is not only caused by changes in the states of chromatin accessibility, but involves other regulatory mechanisms. Compared to the latent representations identified from other prior relational graphs that embed the chromatin accessibility and gene expression data into a shared latent space or into two independent latent spaces, our model can highlight clusters of interest based on the biological

hypothesis of interactions between modalities (Appendix Section E Figure S3a, b). As the conditional latent representation of gene expression aims at modeling the variability that chromatin accessibility data falls short to model, the ability to identify the distinction between cell types using the conditional latent representation is usually limited. By combining latent representations of chromatin accessibility and conditional latent representations of gene expression, our model is also capable of modeling the distinctions between most cell types (Appendix Section E Figure S3c, d).

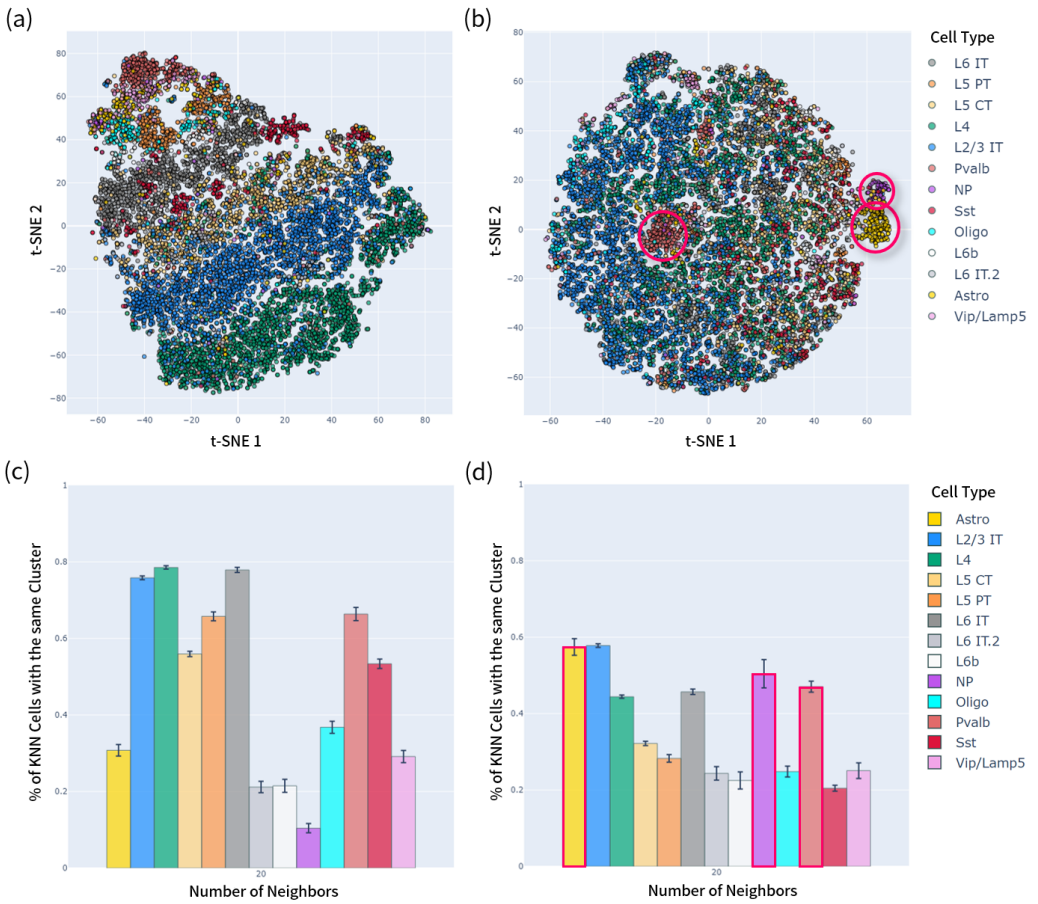

Figure 3: Conditional latent representations for SNARE-Seq cerebral cortex of an adult mouse dataset. (a) The posterior mean $q_\phi^{(atac)}(z_{atac}|x_{atac})$. (b) the posterior mean $q_\phi^{(rna)}(z_{rna}|x_{rna})$, with red circles highlighting the Astro, NP and Pvalb cell types. (c) The proportion of 20-nearest neighbors with the same cell type for the posterior mean $q_\phi^{(atac)}(z_{atac}|x_{atac})$, and (d) the proportion of 20-nearest neighbors with the same cell type for the posterior mean $q_\phi^{(rna)}(z_{rna}|x_{rna})$. The data points are colored by the annotated cell types.

3.2 Unsupervised Multi-facet Clustering

The analysis shown in Method 3.1 can only be achieved when the cell type annotation is available for each cell. When this additional information is not available, *unsupervised multi-facet clustering* can be used to cluster the cells. Note that the clustering strategy for each modality is learned during the training process; therefore, no additional adjustment of the clustering algorithm is required. In addition, *conditional attribution scores* can be used to identify cell clusters of interests. Conditional attribution scores represent the attribution of each dimension in the conditional latent distributions to the generated data. In our proposed framework, conditional attribution scores are computed using integrated gradients [39] while considering the dependencies between latent distributions (Appendix Section G) .

Here, we show that the model is capable of learning clustering assignments conditionally in an unsupervised way for different modalities (Appendix Figure S4). Using the conditional attribution scores, we can highlight the cluster of interests (Figure 4). The clusters with high conditional attribution scores are (a) Cluster 17, which corresponds to Astro cells (F1 score: 0.669, precision: 0.873), (b) Cluster 20, which corresponds to NP cells (F1 score: 0.669, precision: 0.625), (c) Cluster 30, which corresponds to part of the Pvalb cells (F1 score: 0.338, precision: 0.789), and (d) Cluster 19, which corresponds to mixed cell types of the Vimp/Lamp5 (F1 scores: 0.219, precision: 0.344) and L6 IT.2 (F1 scores: 0.183, precision: 0.279). Using the clusters identified with either annotated cell types or unsupervised multi-facet clustering, *Bayesian differential analysis* can be used to identify differentially expressed genes or chromatin accessibility region (Appendix H).

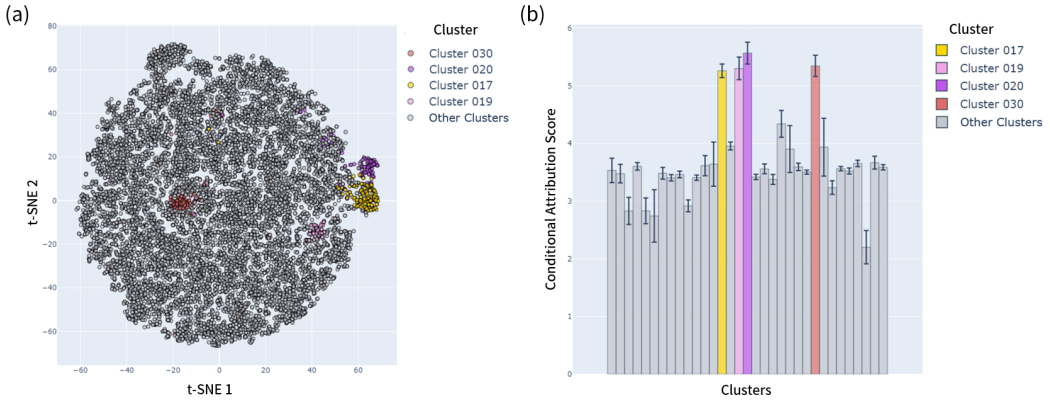

Figure 4: Conditional attribution scores can be used to identify clusters of interests. (a) The posterior mean $q_\phi^{(rna)}(z_{rna}|x_{rna})$. (b) The conditional attribution scores of the latent distribution of gene expression conditioned on chromatin accessibility data to the generated gene expression data. The clusters with high conditional attribution scores are highlighted.

4 Discussion

4.1 Posterior Approximation

Many multi-modal data integration methods applied either the product-of-expert (PoE) [40, 22, 23] $q_\phi(z_\mathcal{V}|x_\mathcal{V}) = \prod_{v\in\mathcal{V}} q_\phi^{(v)}(z_v|x_v)$ or (additive) mixture-of-expert (MoE) [41, 24] factorization to approximate the posterior $q_\phi(z_\mathcal{V}|x_\mathcal{V}) = \sum_{v\in\mathcal{V}} \frac{1}{|\mathcal{V}|} q_\phi^{(v)}(z_v|x_v)$. Here, we mainly discussed the posterior approximation of our created model in the context of PoE. However, as previous studies suggested, MoE factorization might be more suitable for multi-modal data integration [41].

In addition, the Gaussian posterior may limit the expressiveness of the generative model. To address this, we could further exploit the hierarchical structure of variational autoencoder and approximate L layers of posterior distributions for each modality $q_\phi(z_\mathcal{V}|x_\mathcal{V}) = \prod_{v\in\mathcal{V}} \prod_{l=1}^{L} q_\phi^{(v,l)}(z_{v,l}|z_{v,l+1})$, with $z_{v,L} = x_v$. Alternatively, we can incorporate normalizing flows [42, 43] or diffusion models [44] to approximate tighter lower bounds. We leave the exploration and evaluation of different strategies to approximate the posterior for future research and improvement.

4.2 Incorporating Prior Knowledge of Multi-modal Data

Recently, GLUE [26] was proposed to incorporate a prior knowledge graph that represents regulatory interactions between features in different molecular assays to integrate information between modalities. The prior knowledge graph is then used to link latent representations between modalities. Although this approach shows potential for integrative regulatory inference, it may be demanding to construct the prior knowledge graph, since our understanding of the interactions between regulatory elements is often insufficient. Our method complements GLUE as we simply expect the graph which represents the dependency between modalities (e.g. state of chromatin accessibility may influence

gene expression in general, as opposed to specific chromatin region may influence specific gene expression) and can identify clusters of cells that show non-canonical regulatory patterns. This is particularly useful when understanding of a detailed regulatory network is lacking, for instance, in cancer data analysis. Moreover, although we mainly describe our method in the context of single-cell data integration, our method can be applied to more general settings where the provided dependency is based on modality-level (e.g. time dependency in time-series data analysis). In the future, it may be possible to combine prior knowledge on potential gene regulation with multi-modal data relational graphs.

5 Conclusions

We propose a novel probabilistic learning framework that incorporates prior biological understanding of the relationships between different data modalities, creating a hierarchical variational autoencoder explicitly based on the provided relational graph. The main distinction with our proposed method and the previous studies is that the generative process for each modality uses only the latent distribution of the ancestors modality in the graph. In this way, our method is capable of integrating multi-modal data and modeling the generative process of each modality based on the provided conditional properties.

We showed that our method can integrate paired single-cell multi-omics datasets and identify cell clusters of interest. Specifically, it can distinguish between cell types that cannot be explained through observing a single modality (i.e. chromatin accessibility) of the data. To facilitate downstream analysis, our framework provides conditional attribution score for model interpretation, batch correction, online multi-facet clustering, Bayesian differential analysis, and enrichment analysis. In the future, we plan to further characterize the identified cell clusters through differential gene expression analysis. Followed by identifying differentially expressed genes that cannot be modeled with the states of chromatin variability using the conditional attribution score. We foresee that, in the future, our method will be applied to other paired single-cell multi-omics datasets from, for example, 10X Multiome and CITE-Seq.

References

[1] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377–382, 2009.

[2] Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nature methods*, 9(1):72–74, 2012.

[3] Bogdan Budnik, Ezra Levy, Guillaume Harmange, and Nikolai Slavov. Scope-ms: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome biology*, 19(1):1–12, 2018.

[4] Darren A Cusanovich, Riza Daza, Andrew Adey, Hannah A Pliner, Lena Christiansen, Kevin L Gunderson, Frank J Steemers, Cole Trapnell, and Jay Shendure. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237):910–914, 2015.

[5] Blue B Lake, Song Chen, Brandon C Sos, Jean Fan, Gwendolyn E Kaeser, Yun C Yung, Thu E Duong, Derek Gao, Jerold Chun, Peter V Kharchenko, et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nature biotechnology*, 36(1):70–80, 2018.

[6] Anja Mezger, Sandy Klemm, Ishminder Mann, Kara Brower, Alain Mir, Magnolia Bostick, Andrew Farmer, Polly Fordyce, Sten Linnarsson, and William Greenleaf. High-throughput chromatin accessibility profiling at single-cell resolution. *Nature communications*, 9(1):1–6, 2018.

[7] Kevin Grosselin, Adeline Durand, Justine Marsolier, Adeline Poitou, Elisabetta Marangoni, Fariba Nemati, Ahmed Dahmani, Sonia Lameiras, Fabien Reyal, Olivia Frenoy, et al. High-throughput single-cell chip-seq identifies heterogeneity of chromatin states in breast cancer. *Nature genetics*, 51(6):1060–1066, 2019.

[8] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell rna-seq with unique molecular identifiers. *Nature methods*, 11(2):163–166, 2014.

[9] Anna SE Cuomo, Daniel D Seaton, Davis J McCarthy, Iker Martinez, Marc Jan Bonder, Jose Garcia-Bernardo, Shradha Amatya, Pedro Madrigal, Abigail Isaacson, Florian Buettner, et al. Single-cell rna-sequencing of differentiating ips cells reveals dynamic genetic effects on gene expression. *Nature communications*, 11(1):1–14, 2020.

[10] Ansuman T Satpathy, Naresha Saligrama, Jason D Buenrostro, Yuning Wei, Beijing Wu, Adam J Rubin, Jeffrey M Granja, Caleb A Lareau, Rui Li, Yanyan Qi, et al. Transcript-indexed atac-seq for precision immune profiling. *Nature medicine*, 24(5):580–590, 2018.

[11] Silvia Domcke, Andrew J Hill, Riza M Daza, Junyue Cao, Diana R O'Day, Hannah A Pliner, Kimberly A Aldinger, Dmitry Pokholok, Fan Zhang, Jennifer H Milbank, et al. A human cell atlas of fetal chromatin accessibility. *Science*, 370(6518):eaba7612, 2020.

[12] Ricard Argelaguet, Anna SE Cuomo, Oliver Stegle, and John C Marioni. Computational principles and challenges in single-cell data integration. *Nature biotechnology*, 39(10):1202–1215, 2021.

[13] 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. *Nature methods*, 14(9):865–868, 2017.

[14] Stephen J Clark, Ricard Argelaguet, Chantriolnt-Andreas Kapourani, Thomas M Stubbs, Heather J Lee, Celia Alda-Catalinas, Felix Krueger, Guido Sanguinetti, Gavin Kelsey, John C Marioni, et al. scnmt-seq enables joint profiling of chromatin accessibility dna methylation and transcription in single cells. *Nature communications*, 9(1):1–9, 2018.

[15] Song Chen, Blue B Lake, and Kun Zhang. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature biotechnology*, 37(12):1452–1457, 2019.

[16] Sai Ma, Bing Zhang, Lindsay M LaFave, Andrew S Earl, Zachary Chiang, Yan Hu, Jiarui Ding, Alison Brack, Vinay K Kartha, Tristan Tay, et al. Chromatin potential identified by shared single-cell profiling of rna and chromatin. *Cell*, 183(4):1103–1116, 2020.

[17] Kamila Belhocine, Laura DeMare, and Olivia Habern. Single-cell multiomics: Simultaneous epigenetic and transcriptional profiling: 10x genomics shares experimental planning and sample preparation tips for the chromium single cell multiome atac+ gene expression system. *Genetic Engineering & Biotechnology News*, 41(1):66–68, 2021.

[18] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck III, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.

[19] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12):1289–1296, 2019.

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

[21] Danila Bredikhin, Ilia Kats, and Oliver Stegle. Muon: multimodal omics analysis framework. *Genome Biology*, 23(1):1–12, 2022.

[22] Chunman Zuo and Luonan Chen. Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data. *Briefings in Bioinformatics*, 22(4):bbaa287, 2021.

[23] Boying Gong, Yun Zhou, and Elizabeth Purdom. Cobolt: integrative analysis of multimodal single-cell sequencing data. *Genome biology*, 22(1):1–21, 2021.

[24] Kodai Minoura, Ko Abe, Hyunha Nam, Hiroyoshi Nishikawa, and Teppei Shimamura. A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. *Cell reports methods*, 1(5):100071, 2021.

[25] 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. *Nature methods*, 18(3):272–282, 2021.

[26] Zhi-Jie Cao and Ge Gao. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, pages 1–9, 2022.

[27] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Learning hierarchical features from generative models. *arXiv preprint arXiv:1702.08396*, 2017.

[28] Sandy L Klemm, Zohar Shipony, and William J Greenleaf. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, 20(4):207–220, 2019.

[29] Fabian Falck, Haoting Zhang, Matthew Willetts, George Nicholson, Christopher Yau, and Chris C Holmes. Multi-facet clustering variational autoencoders. *Advances in Neural Information Processing Systems*, 34:8676–8690, 2021.

[30] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

[31] Tal Ashuach, Daniel A Reidenbach, Adam Gayoso, and Nir Yosef. Peakvi: A deep generative model for single-cell chromatin accessibility analysis. *Cell reports methods*, 2(3):100182, 2022.

[32] Zhiyuan Li, Jaideep Vitthal Murkute, Prashnna Kumar Gyawali, and Linwei Wang. Progressive learning and disentanglement of hierarchical representations. *arXiv preprint arXiv:2002.10549*, 2020.

[33] Bohan Wu, Suraj Nair, Roberto Martin-Martin, Li Fei-Fei, and Chelsea Finn. Greedy hierarchical variational autoencoders for large-scale video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2318–2328, 2021.

[34] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[35] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: An unsupervised and generative approach to clustering. *arXiv preprint arXiv:1611.05148*, 2016.

[36] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

[37] Matthew Willetts and Brooks Paige. I don't need u: Identifiable non-linear ica without side information. *arXiv preprint arXiv:2106.05238*, 2021.

[38] Bosiljka Tasic, Zizhen Yao, Lucas T Graybuck, Kimberly A Smith, Thuc Nghi Nguyen, Darren Bertagnolli, Jeff Goldy, Emma Garren, Michael N Economo, Sarada Viswanathan, et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, 563(7729):72–78, 2018.

[39] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[40] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in Neural Information Processing Systems*, 31, 2018.

[41] Yuge Shi, Brooks Paige, Philip Torr, et al. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in Neural Information Processing Systems*, 32, 2019.

[42] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[43] Jianlin Su and Guang Wu. f-vaes: Improve vaes with conditional flows. *arXiv preprint arXiv:1809.05861*, 2018.

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[45] Isaac Virshup, Sergei Rybakov, Fabian J Theis, Philipp Angerer, and F Alexander Wolf. anndata: Annotated data. *bioRxiv*, 2021.

[46] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[47] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Swish: a self-gated activation function. *arXiv preprint arXiv:1710.05941*, 7(1):5, 2017.

[48] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[49] External RNA Controls Consortium Lreid@ expressionanalysis. com. Proposed methods for testing and selecting the ercc external rna controls. *BMC genomics*, 6(1):150, 2005.

[50] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[51] Zhuo-Qing Fang. Gene set enrichment analysis in python. `https://github.com/zqfang/GSEApy`, 2022.

[52] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

Appendix

A Decomposition of ELBO

Consider M multiple modalities of data with the same anchor and a directed acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_m \mid m = 1, 2...M\}$ is the set of vertices representing the data modalities and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ the set of unweighted directed edges specifying prior knowledge on the conditional independent relationships between the different modalities. We denote a directed walk from vertex v_i to v_j as $w(i, j)$ and the set of all directed walks in the graph as $\mathcal{W}$. The set of ancestors of a vertex j can then be defined as $\mathcal{A}(j) = \{i | w(i, j) \in \mathcal{W}\}$. The evidence lower bound (ELBO) of the conditional data likelihood can be derived as follows:

$$\begin{aligned}
\log p(x_\mathcal{V}|b) &= \mathbb{E}_{q_\phi(z_\mathcal{V}, c_\mathcal{V}|x_\mathcal{V}, b)}[\log(\frac{p_\theta(x_\mathcal{V}, z_\mathcal{V}, c_\mathcal{V}|b)}{q_\phi(z_\mathcal{V}, c_\mathcal{V}|x_\mathcal{V}, b)})] + \mathbb{E}_{q_\phi(z_\mathcal{V}, c_\mathcal{V}|x_\mathcal{V}, b)}[\log(\frac{q_\phi(z_\mathcal{V}, c_\mathcal{V}|x_\mathcal{V}, b)}{p_\theta(z_\mathcal{V}, c_\mathcal{V}|x_\mathcal{V}, b)})] \\
&= \mathbb{E}_{q_\phi(z_\mathcal{V}, c_\mathcal{V}|x_\mathcal{V}, b)}[\log(\frac{p_\theta(x_\mathcal{V}, z_\mathcal{V}, c_\mathcal{V}|b)}{q_\phi(z_\mathcal{V}, c_\mathcal{V}|x_\mathcal{V}, b)})] + D_{\mathrm{KL}}(q_\phi(z_\mathcal{V}, c_\mathcal{V}|x_\mathcal{V}, b)||p_\theta(z_\mathcal{V}, c_\mathcal{V}|x_\mathcal{V}, b)) \\
&\geq \mathbb{E}_{q_\phi(z_\mathcal{V}, c_\mathcal{V}|x_\mathcal{V}, b)}[\log(\frac{p_\theta(x_\mathcal{V}, z_\mathcal{V}, c_\mathcal{V}|b)}{q_\phi(z_\mathcal{V}, c_\mathcal{V}|x_\mathcal{V}, b)})] = \mathcal{L}(x_\mathcal{V}, b; \theta, \phi)
\end{aligned} \tag{S1}$$

, where $q_\phi(z_\mathcal{V}, c_\mathcal{V}|x_\mathcal{V}, b)$ are parameterized by encoder neural networks $f_\mathcal{V}^{(e)}$. Assuming independence for priors across modalities, the generative model (Equation 1) can be structured as:

$$\begin{aligned}
p_\theta(x_\mathcal{V}, z_\mathcal{V}, c_\mathcal{V}, b) &= p_\theta(b) \prod_{v \in \mathcal{V}} p_\theta^{(v)}(x_v|z_v, z_{\mathcal{A}(v)}, b) p_\theta^{(v)}(z_v|c_v) p_\theta^{(v)}(c_v) \\
&= p_\theta(b) \prod_{v \in \mathcal{V}} p_\theta^{(v)}(x_v|z_v, z_{\mathcal{A}(v)}, b) p_\theta^{(v)}(c_v|z_v) p_\theta^{(v)}(z_v) \\
p_\theta(x_\mathcal{V}, z_\mathcal{V}, c_\mathcal{V}|b) &= \prod_{v \in \mathcal{V}} p_\theta^{(v)}(x_v|z_v, z_{\mathcal{A}(v)}, b) p_\theta^{(v)}(c_v|z_v) p_\theta^{(v)}(z_v)
\end{aligned} \tag{S2}$$

Assuming conditional independence between the posterior of the latent representation and cluster assignment given $x_\mathcal{V}$ and b, the the posterior distribution can be further decomposed:

$$\begin{aligned}
q_\phi(z_\mathcal{V}, c_\mathcal{V}|x_\mathcal{V}, b) &= q_\phi(z_\mathcal{V}|x_\mathcal{V}, b) q_\phi(c_\mathcal{V}|x_\mathcal{V}, b) \\
&= \prod_{v \in \mathcal{V}} q_\phi^{(v)}(z_v|x_v, b) q_\phi^{(v)}(c_v|x_v, b)
\end{aligned} \tag{S3}$$

. Substituting Equation S2 and S3 back to Equation (S1), we can rewrite the ELBO as:

$$\begin{aligned}
\mathbb{E}_{q_\phi(z_\mathcal{V}, c_\mathcal{V}|x_\mathcal{V}, b)} =& [\log(\frac{\prod_{v \in \mathcal{V}} p_\theta^{(v)}(x_v|z_v, z_{\mathcal{A}(v)}, b) p_\theta^{(v)}(c_v|z_v) p_\theta^{(v)}(z_v)}{\prod_{v \in \mathcal{V}} q_\phi^{(v)}(z_v|x_v, b) q_\phi^{(v)}(c_v|x_v, b)})] \\
=& \mathbb{E}_{q_\phi(z_\mathcal{V}, c_\mathcal{V}|x_\mathcal{V}, b)}[\sum_{v \in \mathcal{V}} \log(p_\theta^{(v)}(x_v|z_v, z_{\mathcal{A}(v)}, b))] - \\
& \mathbb{E}_{q_\phi(c_\mathcal{V}|x_\mathcal{V}, b)} D_{KL}(q_\phi(z_\mathcal{V}|x_\mathcal{V}, b)||p_\theta(z_\mathcal{V})) - \\
& \mathbb{E}_{q_\phi(z_\mathcal{V}|x_\mathcal{V}, b)} D_{KL}(q_\phi(c_\mathcal{V}|x_\mathcal{V}, b)||p_\theta(c_\mathcal{V}|z_\mathcal{V}))
\end{aligned} \tag{S4}$$

B Multi-facet Clustering

To identify the optimal cluster assignment for the posterior distribution $q_\phi(c_\mathcal{V}|x_\mathcal{V}, b)$, we applied the single-facet VaDE trick [29, 35] to each vertex to derive the optimal posterior approximation $q_\phi^*(c_\mathcal{V}|x_\mathcal{V}, b)$ that minimizes $D_{KL}(q_\phi(c_\mathcal{V}|x_\mathcal{V}, b)||p_\theta(c_\mathcal{V}|z_\mathcal{V}))$:

$$\underset{q_\phi(c_\mathcal{V}|x_\mathcal{V},b)}{\operatorname{argmin}} D_{KL}(q_\phi(c_\mathcal{V}|x_\mathcal{V},b)||p_\theta(c_\mathcal{V}|z_\mathcal{V})) = \frac{\exp\left(\mathbb{E}_{q_\phi(z_\mathcal{V}|x_\mathcal{V},b)}[\log p(c_\mathcal{V}|z_\mathcal{V})]\right)}{Z(q_\phi(z_\mathcal{V}|x_\mathcal{V},b))}$$
$$, Z(q_\phi(z_\mathcal{V}|x_\mathcal{V},b)) = \sum_{k=1}^{K} \exp\left(\mathbb{E}_{q_\phi(z_\mathcal{V}|x_\mathcal{V},b)}[\sum_{v\in\mathcal{V}} \log p(c_v = k|z_v)]\right) \tag{S5}$$

, where K is the number of components in the mixture of independent Gaussian priors.

C Implementation Details

The input and output formats of our proposed framework are compatible with widely used packages in the field of computational biology, such as AnnData [45], Scanpy and MUON (Figure S1). The created graphical model is implemented using Tensorflow 2.8.0 sequential and functional APIs [46]. Each data modality is handled by a component module with custom data preprocessor, encoder $f_v^{(e)}$, hierarchical encoder ($f_v^{(r)}$ and $f_v^{(b)}$) and decoder $f_v^{(d)}$ (Figure S2). The framework also allows the user to use one component that takes multiple modalities as input, so that the component can also be used as a standalone multi-modal variational autoencoder with a shared latent space across modalities (see Figure 1a).

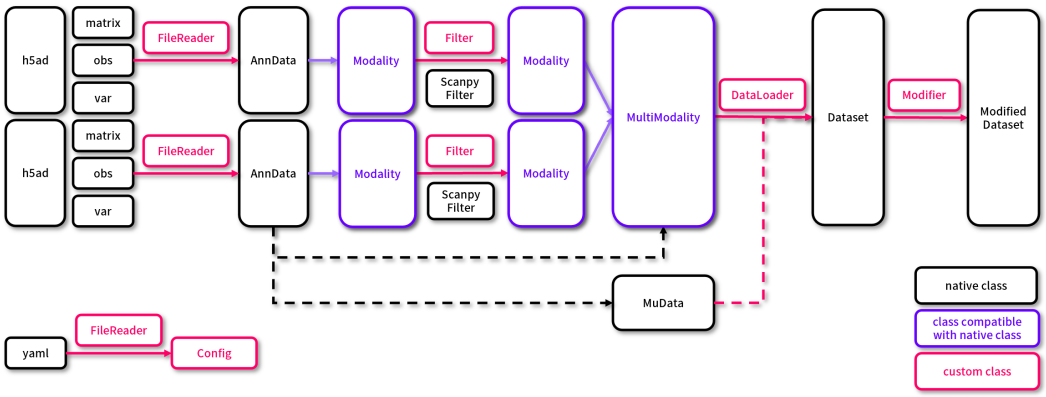

Figure S1: Data flow with the our proposed framework. The framework takes data matrix, cell annotation, and feature annotation (or alternatively h5ad files) as inputs, constructs Modality objects (which is compatible with AnnData), and merges these into a MultiModality object (which is compatible with MuData). Finally, a Tensorflow Dataset used to load the data into the model is created from the MultiModality object.

Note that the decoder of view $f_v^{(d)}$ takes the sampling values from the latent representations of its parents $\hat{z}_{\mathcal{P}(v)}$ as inputs, where $\hat{z}_{\mathcal{P}(v)}$ also incorporates the latent representations of their parents $\bigcup_{i\in\mathcal{P}(v)} \mathcal{P}(i)$ (see Equation C). These dependencies may propagate back to the root (i.e. the modalities without parent). In other words, the posterior $q_\phi(z_v, c_v|x_v, b)$ would be conditioned on the latent representations of all its ancestors. To make the framework more flexible, we also allow the posterior of a specific modality to be conditioned only on the latent representations of its parents $z_{\mathcal{P}(v)}$ instead of all the ancestors $z_{\mathcal{A}(v)}$, using the $AlternativeIncorporateParents$ function:

$$AlternativeIncorporateParents(m) = f_m^{(b)}([z_\mathcal{P}; \tilde{z}_m]; \theta_m^{(b)})$$

To facilitate the future development and broaden the applications of this framework, we further aim to provide high-level APIs as command line tools for researchers, mid-level APIs that allow programmers to customize the model with parameters in the constructor functions, and low-level APIs for developers who wish to design their own data distribution, preprocessing steps, loss function or architecture of the neural network. The default neural network architecture used in the analysis presented in this study is shown in Table S1.

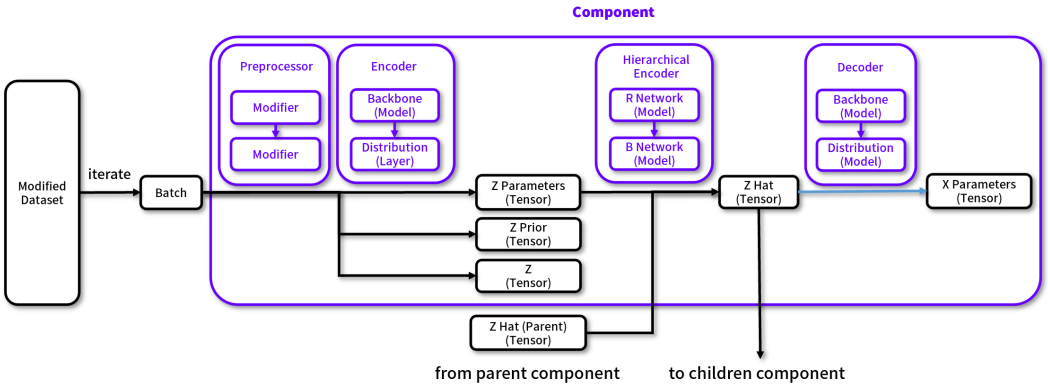

Figure S2: Modules in each component. Every base module, including Preprocessor, Encoder, Hierarchical Encoder and Decoder, can be inherited and extended to create custom modules.

Table S1: Default neural network architecture

Name	Shape of Trainable Weights	Activation Function
$f_m^{(e)}$	$\dim(x_m) \times 1024$	linear
	1024×512	Swish [47], layer normalization [48]
	512×256	Swish, layer normalization
	256×128	Swish, layer normalization
	$128 \times (2 \times \dim(z_m))$	μ (linear), σ^2 (softplus)
$f_m^{(b)}$	$\dim([\hat{z}_{\mathcal{P}}; \tilde{z}_m]) \times \dim(z_m)$	linear
$f_m^{(r)}$	$\dim(z_m) \times \dim(z_m)$	linear
$f_m^{(d)}$	$\dim([z_m; b]) \times 128$	Swish, layer normalization
	128×256	Swish, layer normalization
	256×512	Swish, layer normalization
	$512 \times (\dim(x_m) \times$ number of parameters$)$	custom [1]

[1] with specific activation functions according to the properties of parameters.

D Filtering Parameters

For the gene expression data, we excluded cells with less than 20 expressed genes. Cells with more than 20% of total read count in genes from the External RNA Control Consortium (ERCC) [49] or mitochondria genes are also excluded. For gene expression data, we only preserved genes that had minimum count of 10 across all cells, and that were expressed in at least five cells. For the chromatin accessibility data, we preserved chromatin regions with non-zero counts in at most 10% of the cells and at least 5 cells. The resulting data consisted of 10,309 cells, 19,259 genes and 225,333 chromatin accessibility regions.

E Independent and Joint Modeling

As a comparative analysis, we applied our framework in two different ways: (1) Embedding the single-cell gene expression data from the SNARE-Seq datasets into latent distributions independently without using chromatin accessibility data (Figure S3a), (2) embedding both single-cell gene expression and chromatin accessibility data into a shared joint latent space (Figure S3b). We can observe that independently embedding the gene expression data can better separate cell types in the latent space. This is expected, as the cell type annotation is merely derived from the gene expression data. Since we use the same number of latent dimensions ($\dim(z) = 20$), considering an additional modality (i.e. chromatin accessibility) may limit the expressiveness of the generative process of gene expression data. While we can observe distinct clusters for each cell types using independent embeddings, we can only conclude that this distinction is because of the expression profiles—the connection between

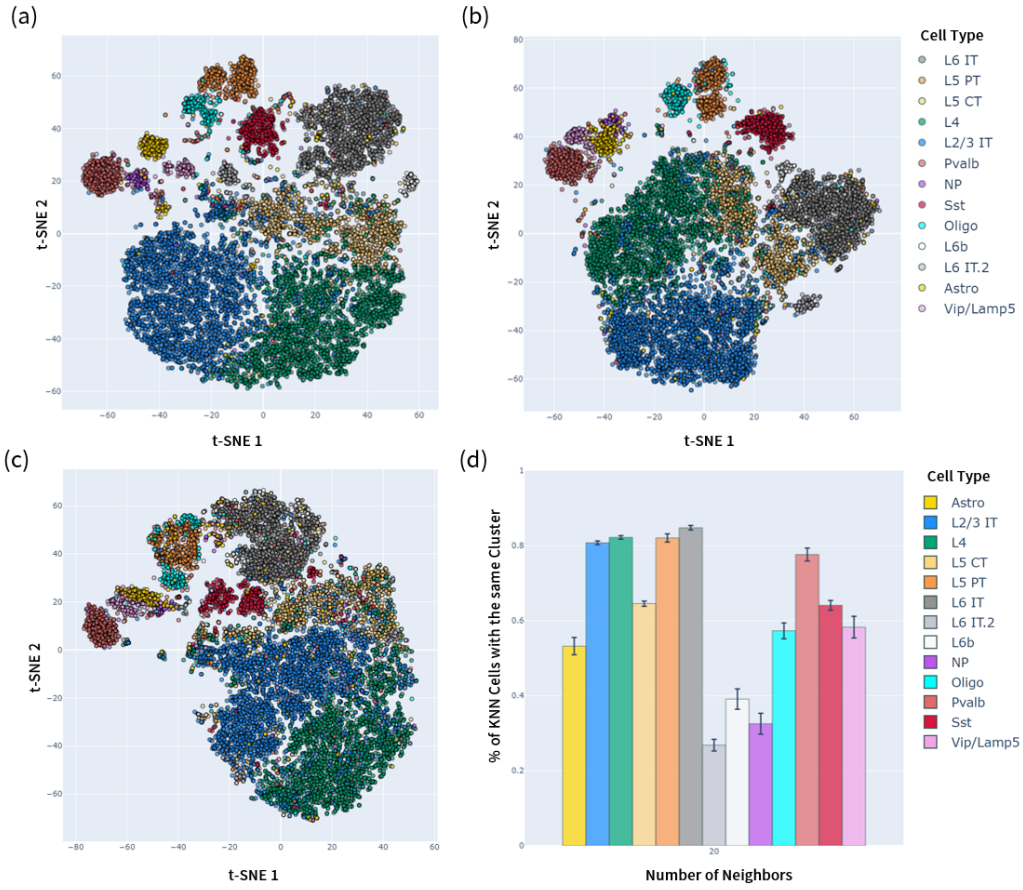

Figure S3: Independent and joint modeling of SNARE-Seq data obtained from cerebral cortex of an adult mouse. (a) The posterior mean $q_\phi(z_{rna}|x_{rna})$ using independent embeddings without considering chromatin accessibility data. (b) The posterior mean $q_\phi(z_{joint}|x_{rna}, x_{atac})$ using joint embeddings. (c) The posterior mean $q_\phi(\hat{z}_{rna}|x_{rna}, x_{atac})$ (not parameterized) using hierarchical embedding. (d) The proportion of 20-nearest neighbors with the same cell type for the posterior mean $q_\phi(\hat{z}_{rna}|x_{rna}, x_{atac})$. The data are colored by the annotated cell types.

differential expression and chromatin accessibility remains unknown. On the other hands, embedding the data into a shared latent space cannot isolate the source of variability between data modalities. As our created model decomposes the generative data distributions of gene expression into two conditional distributions $p_\theta(x_{rna}|\hat{z}_{rna}) = p_\theta(x_{rna}|z_{atac}, z_{rna}) = p_\theta(x_{rna}|z_{rna})p_\theta(x_{rna}|z_{atac})$, it is possible to isolate the source of variability (Figure 3) while modeling the distinction between cell types (Figure S3c, d).

F Unsupervised Multi-facet Clustering

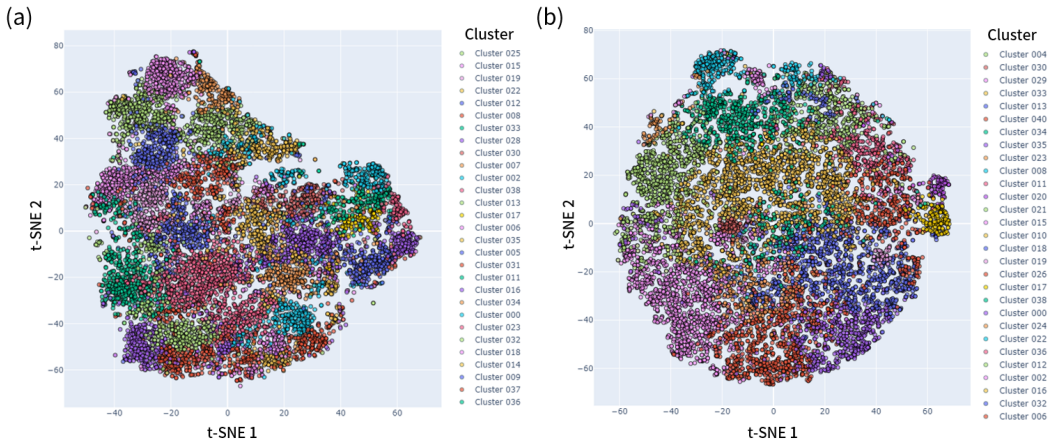

Figure S4: Multi-facet clustering of SNARE-Seq cerebral cortex of an adult mouse dataset. (a) The posterior mean $q_\phi(z_{atac}|x_{atac})$. (b) The posterior mean $q_\phi(z_{rna}|x_{rna})$. The data points are colored by the unsupervised cluster assignments.

G Conditional Attribution Analysis

We use a conditional attribution score with integrated gradient [39] to identify clusters of interests in the conditional settings, which provides better interpretation of the model. The conditional attribution score for modality m is defined as:

$$ConditionalAttributionScore(m) = (\rho_m - \rho_m^{(baseline)}) \int_{\alpha=0}^{1} \frac{\partial \rho_m^{(\alpha)}}{\partial z_m} d\alpha \tag{S6}$$

, where $\rho_m^{(\alpha)} = f_m^{(d)}([f_m^{(b)}([\hat{z}_{\mathcal{P}}; \alpha\tilde{z}_m]; \theta_m^{(b)}), b]; \theta_m^{(d)})$ and $\rho_m^{(baseline)} = \rho_m^{(0)}$. The higher the conditional attribution score, the more relevant the latent distributions of modality z_m is to generate the observed modality x_m. On the contrary, a low conditional attribution score implies that the latent distributions of the ancestors $z_{\mathcal{A}}$ is sufficient to model the generative process of the observed modality x_m. By computing the expected conditional attribution score for each cluster, it allows us to identify clusters of interests in an unsupervised manner.

H Bayesian Differential Analysis

Finally, we implemented an approach to perform differential analysis between the identified clusters. To identify differentially expressed genes or chromatin regions, we compute the Bayesian factor of two hypotheses $\mathcal{H}_1$ and $\mathcal{H}_2$ (see Equation S7) following previous research [30]. Specifically, consider all possible batches $\mathcal{B}$, two groups of anchors $\mathcal{I}, \mathcal{J}$ and the pairs of anchors $(i, j) \subseteq \mathcal{I} \times \mathcal{J}$, where $i \in \mathcal{I}, j \in \mathcal{J}$. The Bayesian factor K to evaluate whether the modality x_m is differentially expressed between group $\mathcal{I}$ and $\mathcal{J}$ is computed as:

$$\begin{aligned} \mathcal{H}_1 &: \mathbb{E}_{\mathcal{B}}[p_\theta(x_m^{(i)}|z_{\mathcal{B}}^{(i)}, b_m^{(i)})] > \mathbb{E}_{\mathcal{B}}[p_\theta(x_m^{(j)}|z_{\mathcal{B}}^{(j)}, b_m^{(j)})] \\ \mathcal{H}_2 &: \mathbb{E}_{\mathcal{B}}[p_\theta(x_m^{(i)}|z_{\mathcal{B}}^{(i)}, b_m^{(i)})] \leq \mathbb{E}_{\mathcal{B}}[p_\theta(x_m^{(j)}|z_{\mathcal{B}}^{(j)}, b_m^{(j)})] \\ K &= \log \frac{p(\mathcal{H}_1|x_m^{(i)}, x_m^{(j)})}{p(\mathcal{H}_2|x_m^{(i)}, x_m^{(j)})} \end{aligned} \tag{S7}$$

. To derive the Bayesian factor K, we use naive Monte Carlo sampling to approximate $\mathbb{E}_{\mathcal{B}}[p_\theta(x_m^{(i)}|z_{\mathcal{B}}^{(i)}, b_m^{(i)})]$ and $\mathbb{E}_{\mathcal{B}}[p_\theta(x_m^{(j)}|z_{\mathcal{B}}^{(j)}, b_m^{(j)})]$. As shown in previous research [30], Bayesian

differential analysis allows our model to identify genes (or chromatin regions) that are consistently differentially expressed across batches. Using the identified differentially expressed genes (or chromatin regions), gene set enrichment analysis (GSEA) [50], can be performed with integrated tools to identify enriched pathways or biological functions [51]. The differentially expressed genes and enriched pathways between Pvalb cells and other cells are shown in Figure S5.

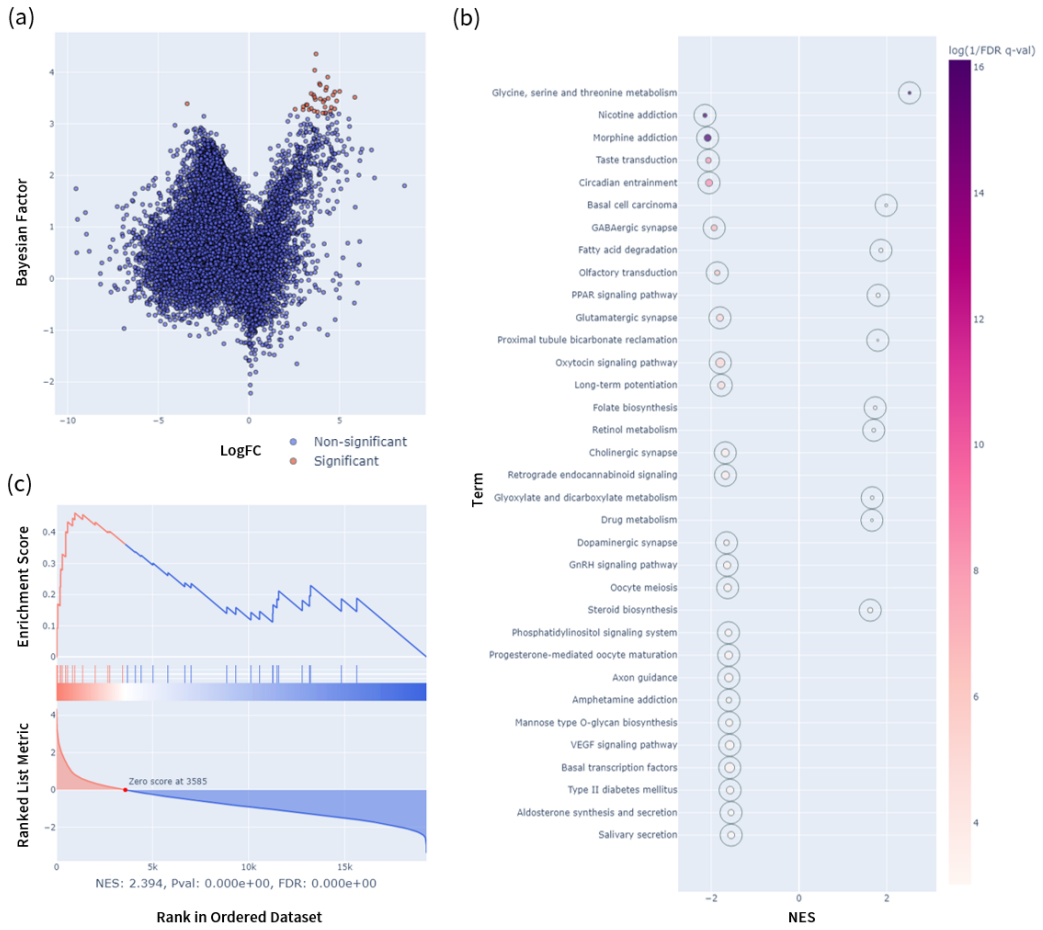

Figure S5: (a) Volcano plot depicting log fold change (logFC) and Bayesian Factors from the differentially expressed gene analysis comparing Pvalb cells against other available cell types. Significant genes are highlighted with red. (b) Ringplot illustrating enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [52] obtained from applying GSEA on the ranked Bayesian factors. NES: Normalized Enrichment Score from GSEA. (c) Enrichment scores for genes in the most enriched pathway—"glycine, serine and threonine metabolism".