Journal Title Here, 2025, pp. 1–31

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

Unifying Multimodal Single-Cell Data Using a Mixture of Experts β-Variational Autoencoder-Based Framework

Andrew J. Ashford,[1,*] Trevor Enright,[2, 3] Olga Nikolova[3] and Emek Demir[1, 2, 3]

[1]Department of Molecular and Medical Genetics, Oregon Health & Science University, 3181 Sam Jackson Park Rd., 97239, OR, USA, [2]Computational Biology Department, Brenden-Colson Center for Pancreatic Care, Oregon Health & Science University, 3181 SW Sam Jackson Park Rd., 97239, OR, USA and [3]Cancer Early Detection Advanced Research Center (CEDAR), OHSU Knight Cancer Institute, School of Medicine, Oregon Health & Science University, 2720 S Moody Ave., 97201, OR, USA
*Corresponding author. ashforda@ohsu.edu
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Since 2014, 47 technologies have been developed to measure multiple biological modalities from the same cells. However, tools for robustly analyzing these data to uncover holistic biological interactions remain limited. Advancing this field could transform research for many disciplines, including human disease and cancer. To address this limitation, we present UniVI (**Uni**fied **V**ariational **I**nference), a generalizable deep learning algorithm that aligns single-cell measurements from disparate modalities using β—and mixture-of-experts—variational autoencoder frameworks. UniVI learns a latent embedding for each modality while minimizing the divergence between them, a concept often referred to as manifold alignment. Once trained, UniVI enables batch correction, latent factorization, cell-cell alignment, data denoising, and imputation. We demonstrate its performance on multimodal single-cell datasets, including CITE-seq and 10x Multiome data, showing UniVI outperforms widely-used methods without reliance on prior knowledge. This flexibility and generalizability allows UniVI to adapt to emerging multimodal technologies. Our results highlight UniVI's ability to integrate diverse multimodal and unimodal data, offering a scalable solution for refining biological insights. The unified latent spaces it generates enable exploration of cross-modality correlations and the generation of realistic new data, paving the way for novel discoveries in single-cell biology.

Key words: single-cell, multi-omics, multimodal, machine learning, data integration, variational autoencoder, VAE

Introduction

Over the last decade, developments in sequencing and imaging technologies have led to a cascade of new single-cell measurement techniques. Several of these new techniques can jointly measure multiple modalities. For instance, one method was conceived that can measure joint mRNA abundance and DNA alterations [1], while another was developed around the same time that can measure a combination of DNA methylation levels and mRNA abundances [2]. By 2020, there were nine double-modality sequencing techniques, three triple-modality sequencing techniques, and one quad-modality single-cell sequencing technique [3]. A recent review [4] highlights a total of 47 different multimodal techniques, illustrating the breakneck rate at which these approaches are being developed. Yet, compared to their single-modality counterparts, these joint modalities are less streamlined, more expensive, and come with substantial setup overhead. Additionally, they cannot be easily combined to move to higher modality dimensions.

With the explosion of these technologies comes a major need for the development of analysis tools that can make the most of the measurements they produce. To this end, **there is a critical need to generate workflows that align parallel samples from multimodal data to detect cross-modality correlations**. Can multiple, experimentally disjoint single-cell modalities be aligned to each other computationally? What methods can we use to best suit this task and allow researchers to investigate the richness inherent to complex biological systems within—and across—sequencing modalities?

These measured modalities can be conceptualized as distinct views derived from a shared cellular state space—a representation of the possible cellular states within a sample that the modalities, or views, share [5]. Since the samples originate from the same underlying pool of cells, their frequency distributions along this state space should theoretically be consistent across modalities. This suggests that identifying approximately isometric latent embeddings for each modality could facilitate successful alignment of cells between modalities,

© The Author 2025. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

provided the latent embeddings preserve meaningful structural information and are asymmetrical across all dimensions [6, 7].

By mapping these distinct views back into a common aligned latent space, it becomes feasible to approximate which cells in one modality correspond to those in another. This approach effectively emulates a joint multimodal measurement and enables the integration of information across modalities to uncover shared biological insights [8, 9]. Such computational alignments are foundational for achieving integrated analyses of single-cell datasets, particularly in cases where multimodal measurements are unavailable or experimentally impractical.

In recent years, several general statistical and machine learning approaches have been leveraged to address this problem: anchored alignment [7], canonical correlation analysis (CCA) [10], optimal transport (OT) [11], and more recently deep learning—specifically coupled variational autoencoders (VAEs) for aligning latent manifolds [12, 13, 14, 15, 16]. As with any analysis technique, these approaches all have their own specific strengths and weaknesses. Here, we provide a brief review of these general approaches and highlight the utility of each method, then expand on our work.

Shared feature—or bridged alignment—approaches utilize shared components between modalities to anchor their integration. These can come in two forms: (i) features that measure biologically the same or related entities, e.g. gene expression in single-cell RNA-sequencing (scRNA-seq) and corresponding Multiplexed Error-Robust Fluorescence In Situ Hybridization (MERFISH) probes [17] or (ii) *a priori* assignment of aligned centroids of the common cell states e.g. T-cells. Although some of the earliest approaches were in this general category, they both require substantial experimental design, injection of prior information and in most cases, manual adjustments.

CCA relies on inferring shared information from cross-covariance matrices of features between data modalities [18]. For example, given two vectors X and Y, CCA will infer linear combinations of features in X and Y that have the maximum correlation between one another across datasets. CCA has been used for a wide range of cross-modal correlation problems in biological research, imaging data, and beyond. A number of multimodal single-cell integration workflows have been designed using CCA as a foundational framework [8, 19, 20, 21, 22, 23, 24, 25]. However, CCA's underlying linear assumption is likely too simple for cross-modal relationships in cell-derived data.

OT is formulated as an optimization problem where we find the "optimal path" between two distributions that minimizes an energy cost. There have been numerous applications of this general idea, including biological applications such as cell trajectory inference and multimodal sequencing data integration [26, 27, 28, 29, 30]. However, this approach does not have a common frequency distribution constraint and we risk losing critically important inter—and intra—modal information when deriving cross-distribution calculations.

Existing deep learning-based approaches [12, 13, 14, 15, 16] learn latent embeddings for each modality and align them during training with varying architectures and training approaches. These methods hold the most promise for this particular open biological problem because they are able to capture the complex, non-linear nuances between biological data of multiple modalities. Additionally, these methods eliminate much of the necessity for prior knowledge for a given biological study system and its associated biases [31]. That said, the landscape of approaches that leverage this technique vary significantly when it comes to performance, architecture, method debugging and upkeep, preprocessing steps, ease of use, prior knowledge requirements, and assumptions made for the data used [32]. Additionally, the reach of these tools can be limited given the computational and programming knowledge required to use many of them.

Based on these observations, we developed UniVI, a generative mixture-of-experts (MoE) β-VAE framework [33, 34, 35] (refer to **Figure 1a** for a model architecture schematic). UniVI minimizes symmetric cross-entropy loss [36] between latent spaces for high-precision alignment of single-cell modalities. Our overall strategy is to learn a latent embedding for each modality while minimizing the divergence between them, a concept often referred to as "manifold alignment" [37]. Once trained, these manifolds can offer rich, information-dense opportunities for exploration.

A common technique, for example, is to visualize the learned manifolds in two or three dimensions using dimensionality reduction techniques such as uniform manifold approximation and projection (UMAP) [38] and principal component analysis (PCA) [39, 40] to gain insight into underlying feature distribution structures [32]. While visualizing the latent space can offer a slough of valuable insights and allow researchers to gain a better understanding of the relationships within—and between—cellular populations, it's just the tip of the iceberg when it comes to the potential applications of these information-dense learned structures. For example, we show in this manuscript that once the model learns the underlying manifolds, researchers can then perform several tasks including batch correction, latent factorization, cell-cell alignment, denoising, imputation, and the generation of realistic-looking whole samples from existing data measurements (see **Figure 1b-c** for a schematic overview of use-case examples).

In preliminary studies, our group has successfully used UniVI to map the combined latent space embeddings of multimodal single-cell data using a subset of different modalities, including cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) [41] data in addition to jointly measured scRNA-seq and transposase-accessible chromatin sequencing (scATAC-seq) [42] data—a multimodal sequencing technique developed by 10x Genomics called 10x Multiome [43, 44].

This generalizability gives UniVI the unique ability to accommodate data integration for new technologies as they are developed, including the ability to upscale from dual-modality sequencing methods. Additionally, UniVI is an unsupervised *ab initio* approach which allows it to effectively circumvent the constraints posed by methods that rely on extensive prior knowledge and/or labels. Our analysis shows that UniVI outperforms existing widely-used approaches without reliance on prior knowledge across a broad spectrum of multimodal data techniques.

Materials and Methods

We conducted a series of experiments to evaluate the UniVI multimodal integration workflow. Two primary multimodal data types were utilized for model training and interrogation with a total of three multimodal datasets: a robust CITE-seq dataset [45, 46, 47] and two different 10x Genomics Multiome datasets [43, 44]. Additionally, we performed an experiment to integrate unimodal datasets of different modalities, we also utilized a unimodal scRNA-seq [48] and scATAC-seq [49]

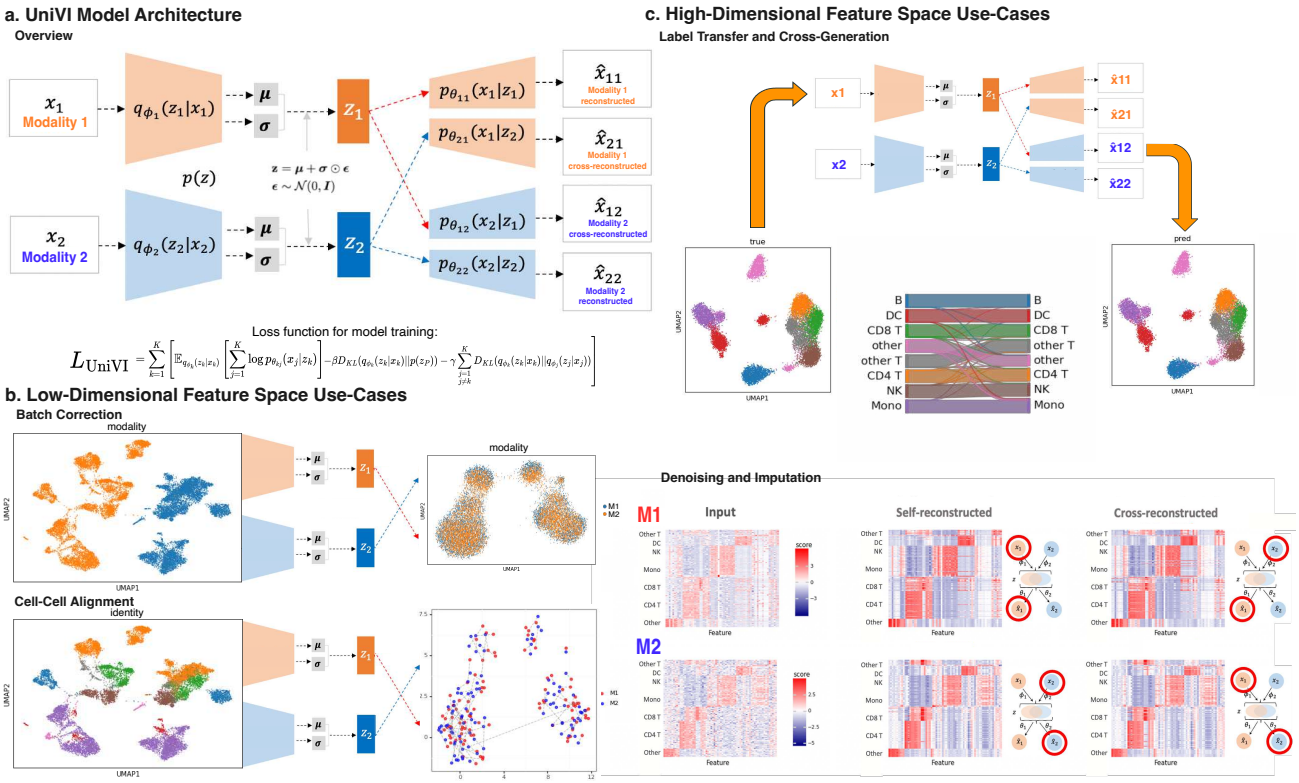

Fig. 1. Overview of the UniVI Model Architecture and Qualitative Use Cases. Here, we highlight schematic examples of different possible use-cases of UniVI. **(a.)** Overview of UniVI model architecture and training loss function. (*Top*) We illustrate the key components of the UniVI model in a dual-modality setup. This includes two total encoders and corresponding latent feature vectors, one of each for each modality. Additionally, there are four decoders—two for each modality, one for self-reconstruction and one for cross-reconstruction. For each component, we include the generalized mathematical representation. (*Bottom*) We depict the loss function employed during model training, emphasizing the objectives of minimizing reconstruction error and maximizing both cross-modal correlation fidelity and latent feature prior distribution similarity. **(b.)** Low-dimensional feature space use-cases—batch correction and cell-cell alignment. (*Top*) We highlight representative multimodal data colored by modality prior to—and after—integration, illustrating the corresponding model components responsible for each modality. (*Bottom*) We further highlight that we can use our method for cell-cell alignment between modalities in the latent space. The data is represented prior to integration colored by cell type. After integration, the alignments of each cell are represented in the latent space using dotted-lines between paired samples from each modality, where red represents modality 1 and blue represents modality 2. **(c.)** High-dimensional feature space applications (i.e. the gene feature space)—data generation, imputation, and annotation transfer. (*Top*) Next, we show a schematic overview of the flow of information when using a trained UniVI model to both transfer labels and generate realistic cross-modal data across modalities. A Sankey plot compares actual annotations with those generated by the model, demonstrating annotation transfer efficacy. (*Bottom*) We highlight representative heatmaps showcasing UniVI's denoising and imputation capabilities. (*Bottom-Left*) Heatmaps of each modality's high-dimensional data before model integration—noting that the features show significant noise. (*Bottom-Middle*) We show a representation of the denoising effect when using our method to reconstruct data unimodally. (*Bottom-Right*) We additionally see this effect when reconstructing data cross-modally, which is highlighted here. Using these learned biologically relevant features, we can also generate—or impute—realistic biological data samples.

dataset. We describe the data in more detail below, highlighting dataset specifics, preprocessing, and feature selection steps.

CITE-Seq Data (Dataset 1)

For all analyses that used CITE-seq data, we used a robust dataset comprising 161,764 PBMCs from eight HIV vaccine trial donors with ages spanning from 20-49 years [45, 46, 47], available here through the Fred Hutchinson Cancer Research Center's dataset portal [50]. Additionally, the raw data can be accessed in the Gene Expression Omnibus (GEO) Database [51] under accession number GSE164378 and all raw sequencing data are available in the database of Genotypes and Phenotypes (dbGaP) [52] under accession number phs002315.v1.p1. The dataset includes 20,729 scRNA-seq-derived total transcript features and 228 total single-cell antibody-derived tags (scADT)-detected surface proteins. Of note, we utilized only the CITE-seq data and left out the ECCITE-seq data for the purposes of our comprehensive analyses in this manuscript.

For this study, we used the Fred Hutchinson Cancer Research Center's h5Seurat file specified above and used a custom R script to convert it into AnnData format [53] for use with Python. This file contains cell type assignments that were originally annotated by Hao et al. [45] utilizing a combination of the scRNA-seq and scADT data and clustering approaches that integrated both modalities. The cell identities were then validated with known markers using both protein and transcript features. Harnessing both modalities for cell type annotation allowed the authors to fine-tune their cell identity assignments to an incredibly high level of detail. These annotations include level1 annotations with eight cell types, level2 annotations with 31 cell types, and level3 annotations with 58 cell types—the highest granularity cell type category [45]. Of note, we performed the same preprocessing steps outlined below for all

the CITE-seq data analysis performed in this manuscript except for the data integration tool performance evaluation, which we outline below.

To avoid class imbalances in the model and since this dataset has a significant number of samples, we used a stratified subsampling method to randomly select a set number of samples per level1 cell type. Additionally, we show that our model can train a high-performing model using a small subset of available data. This was performed using a built-in UniVI function imported from `UniVI.datasets._anndata` called `subsample_anndata()` that randomly selects samples without replacement for a specified `N` samples per specified cell type category. We passed in the following parameters: `N = 1500`, `key_category = 'level1'` with the default random seed `42` for reproducibility. This resulted in 1,500 samples from each of the eight cell types in the level1 category—B cells, CD4+ T cells, CD8+ T cells, dendritic cells, monocytes, natural killer cells, "other" cells, and "other T" cells—for a total of 12,000 cells.

For the integration method performance evaluation, we used all the data available in the CITE-seq dataset for a more robust comparison between methods. We split the whole dataset (161,764 cells) into 20 total subsets and sampled without replacement using the same proportion of each level3 cell type for each subset. This resulted in 20 subsets with 8,088 total cells each, where the cell type representation in each is the same proportion as the full dataset.

For preprocessing, starting with the scRNA-seq modality, we filtered the data by using the built in UniVI `PrepAnnDataRNA()` preprocessing function, which is imported from `UniVI.datasets._anndata`. This function is a wrapper class built on the Scanpy Python module [54] which also harnesses the AnnData module for easy data preprocessing, reproducibility, and object manipulation when using our workflow. To reduce the influence of dying and low quality cells in the data, we removed cells with 200 or less total genes using the `scanpy.pp.filter_cells()` function, we also removed genes with representative counts in three cells or less using the `scanpy.pp.filter_genes()` function. We additionally filtered out cells with a total of 10% or more mitochondrial reads. Next, we filtered 0.5% of outliers which equates to the removal of 0.25% of the data on each extreme of the data distribution. After the initial data preprocessing and feature selection steps outlined above, we selected the top 2,000 transcript features with the highest variability among the remaining data using the `scanpy.pp.highly_variable_genes()` function, which was implemented via our `PrepAnnDataRNA()` wrapper class automatically by passing in the parameter `n_top_genes = 2000`. Following this step, the wrapper class uses Scanpy's `pp.scale()` function to scale the normalized data by transforming the normalized values into Z-scores.

For the scADT modality, we preprocessed the data in a similar manner as we did for the scRNA-seq data in concordance with several manuscripts [55, 56, 57]. For this, we implemented another wrapper class function imported from `UniVI.datasets._anndata` called `PrepAnnDataADT()`. The preprocessing steps for this included the removal of 0.5% of total outliers. Moreover, we chose to remove the four non-human isotype control proteins: "Rat-IgG-1", "Rat-IgG2b", "Rat-IgG1-2", and "Rag-IgG2c"—which brought the total scADT feature count down to 224 from 228. The `PrepAnnDataADT()` wrapper class then performs the Scanpy functions `pp.normalize_total()` to normalize by the total read counts for each feature, `pp.log1p()` to log normalize each feature plus one, then `pp.scale()` to convert the data to Z-scores.

10x Genomics Multiome Data (Datasets 2 and 3)

We used two different 10x Genomics Multiome human PBMC datasets. These datasets both contain joint multimodal scRNA-seq and scATAC-seq measurements of PBMCs from healthy human donors. Of note, they were generated using different 10x Genomics instruments, one utilizing the Chromium X Series instrument (Dataset 2), which can be accessed here [43] and one using the Chromium Controller instrument (Dataset 3), which can be accessed here [44]. These datasets both used the same chemistry and pipeline versions but used samples from different donors. We used the different datasets to show the generalizability of our method and for ease of standardizing features between the 10x Multiome and unimodal datasets.

10x Genomics Multiome Chromium X Series Data (Dataset 2)

This dataset contains multimodal human PBMC measurements from a healthy male donor aged 30-35 [43]. It contains 10,974 recovered PBMC nuclei (out of about 16,100 total transposed nuclei) using the workflow described in the Chromium Next GEM Single-Cell Multiome ATAC + Gene Expression Reagent Kits User Guide (CG000338 Rev E). There were 29,525 total transcripts detected with 13,039 of those being linked to the modality opposite of it. For the scATAC-seq modality, there were 111,742 total peaks detected with 89,426 of them being linked to detected scRNA-seq features. The data were generated using a 10x Genomics Chromium X Series instrument, sequenced on an Illumina NovaSeq 6000, and processed using the cellranger-arc-2.0.0 pipeline version with the Single-Cell Multiome ATAC + Gene Expression v1 chemistry.

Starting with the scRNA-seq modality preprocessing steps, we prepared the raw transcript counts using the previously mentioned in-house `PrepAnnDataRNA()` wrapper function. We trimmed low-quality data by removing cells with 200 or less total gene counts in addition to genes that were only represented in three or fewer cells. Additionally, cells were culled that contained 10% or more mitochondrial transcripts to further address the issue of data resulting from low quality and stressed cells. Next, we removed a total of 0.5% outliers (0.25% from each extreme of the distribution). The wrapper function then normalizes the data by total read counts using Scanpy's `pp.normalize_total()` then log-normalizes the data using the `scanpy.pp.log1p()` function. The data are then scaled and converted to Z-scores using `scanpy.pp.scale()` by setting `scale = True` when calling the function.

To preprocess the scATAC-seq modality data, we used the built-in UniVI `PrepAnnDataATAC()` class to perform several preprocessing steps that closely mirror the `PrepAnnDataRNA()` class steps described above in accordance with several existing studies [56, 57, 58, 59]. The `PrepAnnDataATAC()` class relies on several functions utilizing Scanpy modules for data preprocessing as described above.

Once we filtered, normalized, performed feature selection, and scaled both of the modalities, we concatenated the intersection of the resulting cells. This step resulted in a subset of 9,437 cells with 2,000 scRNA-seq and 5,000 scATAC-seq paired features per cell. We then made two data splits, one for integration model training, validation, and testing and one for downstream analyses to interrogate the trained UniVI models with data unseen to it. For this step, we selected 80% of the samples without replacement from each cell type with 19 total classes. The remainder 20% of each cell type were held

aside as a test set for the cross-modally generated scATAC-seq classification neural network experiment. This ultimately resulted in a model integration set containing 7,541 total cells with 2,000 transcript features and 5,000 scATAC-seq peak features and a separate set of 1,896 total cells with the same feature set held aside.

We separately performed the same process described here to create an 80% integration model set using 2,000 transcript features and 2,000 scATAC-seq peak features. These data splits were used for the scRNA-seq denoising experiment. We selected fewer scATAC-seq features here since we were mostly focusing on the scRNA-seq data denoising for that particular experiment and 2,000 scATAC-seq features were enough to both reduce computational complexity and train well-performing integration models for this use-case.

We observed that preliminary cross-modal classification results performed poorly for minority classes. We hypothesized that this could be due to the model not learning sufficiently deep representations of specific cell types. Since our model includes a Monte Carlo sampling step, we reasoned that even with duplicate samples in the training set, the model still introduces variability by sampling features from the latent space's feature distributions during training. To enhance representation learning, we added a random sampling step, resampling each of the 19 cell types 1,500 times with replacement from the 80% training split. This led to a training set with 28,500 total samples, 1,500 of each class for that specific experiment (Supplemental Figure 8). We included a visualization of this trained models' latent space using the 20% test set of unseen data described above and showed that the sampling step did not cause the model to overfit the data on the unseen samples (Figure 5a).

10x Genomics Multiome Chromium Controller Data (Dataset 3)

This dataset contains multimodal human PBMC measurements from a healthy female donor aged 25 that were obtained by 10x Genomics from AllCells [44]. It contains 12,012 recovered PBMC nuclei as described in the Demonstrated Protocol-Nuclei Isolation for Single-Cell Multiome ATAC + Gene Expression Sequencing (CG000365 Rev A) workflow. There were 30,348 total transcripts detected with 13,958 of those being linked to the features in the scRNA-seq modality. For the scATAC-seq data, there were 111,857 total peaks detected with 79,187 of them being linked to detected transcript features. The data were generated using a 10x Genomics Chromium Controller instrument, sequenced on a Novaseq 6000 v1 Kit (Forward Strand Dual-Index Workflow) and processed using the cellranger-arc-2.0.0 pipeline version with the Single-Cell Multiome ATAC + Gene Expression v1 chemistry.

Since we largely used these data for training a UniVI model to integrate separate unimodal datasets, we needed to standardize the features to the unimodal features of each data modality. We used similar steps outlined in the MultiVI manuscript [60] for data preprocessing of both this dataset and the unimodal single-cell datasets outlined below. To this end, we reprocessed the fragment file using cellranger-arc-2.0.0 with the peaks from Dataset 5 below. We also detected overlapping transcript features between this dataset and Dataset 4 below. After filtering and detecting overlapping features, there were a total of 9,022 transcript features shared between Dataset 3's scRNA-seq modality and Dataset 4, and a total of 85,485 peak features that were shared between Dataset 3's scATAC-seq modality and Dataset 5.

We preprocessed the scRNA-seq and scATAC-seq count data using the same methods described above in the 10x Genomics Multiome Chromium X Series Data (Dataset 2) section. Of note, we dedicated 100% of this dataset's samples to model training, validation, and testing. We reasoned that in order to apply learned features from this dataset to outside data of different modalities, we needed to learn a robust and generalizable feature space. To have the best chance of accomplishing a model capable of this kind of generalization, we needed the model to learn from as many features as possible.

Disjoint Unimodal Data

We sourced two separate unimodal human PBMC datasets to test whether we can integrate disjoint unimodal data using a UniVI model pre-trained using multimodal data. First, we accessed the scRNA-seq dataset from Ding et al. [48], available at GEO: GSE132044 and the Broad Institute Single-Cell Portal [61] with accession numbers SCP424, SCP425 and SCP426. We additionally utilized an scATAC-seq dataset from Satpathy et al. [49], available at GEO: GSE129785. The following subsections detail the data access, preprocessing steps, partitioning strategies, and feature selection approaches.

Ding et al. scRNA-Seq Data (Dataset 4)

This dataset is an scRNA-seq dataset from Ding et al. [48], which leverages several different scRNA-seq sequencing technologies. Specifically, we utilized the hg38 PBMC count, cell, and gene matrices for this analysis. Moreover, we downloaded matching cell type annotation from the Broad Institute SCP. We filtered to remove features that were detected in fewer than 1% of cells. The filtered data contained 30,495 cells and 9,022 transcript features that overlapped with Dataset 3's transcript features. We preprocessed the transcript count data, first by total count normalization using `scanpy.pp.normalize_total()`, then by log1p normalization using `scanpy.pp.log1p()`. Of these transcripts, we filtered down to the same top 2,000 most variable transcript features used for Dataset 3 so both processed datasets contained the same standardized features. The data were then Z-scaled using `scanpy.pp.scale()` with `max_value = 10`. Since we trained the model using the top 2,000 most variable features from Dataset 3, we used the same filtered features described above to generate the latent representations of the data shown in this manuscript.

Satpathy et al. scATAC-Seq Data (Dataset 5)

We sourced a robust scATAC-seq dataset from Satpathy et al. [49]. We filtered the data to include only peaks present in at least 0.1% of the data and used the UCSC liftover utility [62] to transfer these peaks from hg19 to hg38 in order to standardize the features to Dataset 3's peak features. The filtered data contained 47,148 human PBMCs with 85,485 peak feature measurements that overlapped with Dataset 3's peak features. Similarly to Dataset 4 above, we performed the Scanpy `pp.normalize_total()` and `pp.log1p()` functions on the peak count data. We then selected the top 5,000 peak features that were most variable within Dataset 3's scATAC-seq modality. Next, we scaled the data to Z-scores using `scanpy.pp.scale()` with `max_value = 10`. Following model training using both of dataset 3's modalities, the 5,000 shared

scATAC-seq features from this dataset were used to test whether we could successfully integrate Datasets 4 and 5.

UniVI Model Framework

UniVI is a generative VAE-based model built on PyTorch [63] that leverages several recent deep-learning concepts such as MoE and β-VAE frameworks [33, 34, 35] to unify multimodal data. UniVI minimizes symmetric cross-entropy loss [36] between latent spaces for high-precision alignment of single-cell modalities. Our overall strategy is to learn a latent embedding for each modality while minimizing the divergence between them, a concept often referred to as "manifold alignment" [37]. Once trained, these learned manifolds can offer rich, information-dense opportunities for exploration. Here, we give a detailed overview of our method and all its components.

Model Architecture

UniVI is a novel multimodal integration model inspired by the MoE configuration proposed by Shi et al. [35]. It incorporates multiple VAEs [33] to effectively capture complex nonlinear relationships across different data modalities. The architecture was initially inspired as a way to emulate cognitive processes in human multimodal perception, enabling robust integration of heterogeneous data, regardless of modality [35]. Shi et al. use the examples in their manuscript of a human brain integrating sights and sounds to infer information about the same phenomena. Key features that we seek to emulate in the context of biological data with UniVI include:

1. **Latent Factorization:** Efficient disentanglement and factorization of modality-specific latent spaces.
2. **Joint-Modality Correlations:** Coherent generation of joint-modality correlations from continuous latent spaces.
3. **Cross-Modal Generation:** Accurate cross-generation capabilities, reconstructing one modality from another.
4. **Multimodal Synergy Optimization:** Enhanced synergy between modalities, improving overall integration and interpretability.

For dual-modality data, the UniVI architecture incorporates two encoders and two latent spaces, one of each corresponding to each different modality. The latent vectors are then decoded back into the original feature space by a total of four decoders, each of which corresponding to a self reconstruction and a cross reconstruction for each modality (refer to Figure 1a for a graphical representation of this architecture). At its core, UniVI is built on a PyTorch [63] framework—a robust high-performance deep learning library for Python. The architecture is fully customizable, allowing researchers to modify and adjust specific hyperparameters prior to training to ensure the model is tuned to their specific dataset for the best results. This includes the ability to adjust hidden layer and neuron counts for each encoder or decoder, edit the total number of latent dimensions, specify the total high-dimensional feature space for each encoder/decoder pair, and specify the data reconstruction distribution type, to name a few adjustable parameters.

Encoders and Latent Space Representation

The encoders of the model—represented by Equation 1—contain a user-specified number of input features, hidden layers, and nodes per layer which are all adjustable per modality. The latent spaces (z_1 and z_2) are computed using the encoder equation and each consist of a user-defined number of latent dimensions, which we'll call d_{latent}. Each latent dimension represents a probability distribution characterized by a mean and a standard deviation. When combined, the combination of these distributions can be referred to as a manifold [37] in the context of VAEs.

The latent space in a VAE is regularized through the inclusion of a prior distribution and the β term within the loss function, which balances the Kullback-Leibler (KL) divergence [64] between each latent feature and the specified prior. This regularization ensures that the learned latent space adheres to a structured and continuous distribution. Typically, the prior is chosen as a standard normal Gaussian distribution, $p(z_P) = \mathcal{N}(0, I)$, as defined in Equation 2. While alternative priors can be used in model training, $\mathcal{N}(0, I)$ remains the standard choice due to several theoretical and practical advantages. First, it enables the reparameterization trick, which allows efficient backpropagation through stochastic latent variables by expressing the latent representation as a differentiable function of a fixed Gaussian noise term [33]. Second, it facilitates latent space regularization by encouraging smooth and continuous feature representations, ensuring that nearby points in the latent space correspond to semantically similar features. Additionally, the use of a standard normal prior simplifies optimization by providing a closed-form solution for the KL divergence term, reducing computational complexity [33]. Moreover, by constraining the latent space to follow a Gaussian structure, the model avoids overfitting and learns a meaningful, structured representation that allows for efficient sampling and interpolation. This makes it particularly useful for generating realistic samples, as new data points can be obtained by sampling from $\mathcal{N}(0, I)$ without requiring explicit training examples [33]. Collectively, these factors contribute to the effectiveness of VAEs in learning robust, generalizable, and interpretable latent feature spaces.

The latent spaces serve as compact representations that effectively capture the most salient features of the input data, enabling dimensionality reduction, denoising, and data imputation. In our approach, these latent spaces also facilitate the alignment of different data modalities by learning structured manifolds with similar geometric properties. This alignment preserves cross-modal relationships, allowing researchers to systematically uncover shared biological patterns and dependencies across multimodal datasets.

Decoders

Each latent space is connected to two decoders, where each decoder can be generalized with Equation 3. Specifically, it is calculated with either Equation 4a or Equation 4b depending on the distribution of your input data after preprocessing and feature selection steps. Notably, the decoder can be specified to reconstruct the data as either a Gaussian distribution or a negative binomial distribution by setting `params['model_arch_1']['dec_model'] = 'gaussian'` or `params['model_arch_1']['dec_model'] = 'nb'`, respectively.

The decoder shape follows the specified encoder shape. For example, if a user specifies that encoder 1 has 2,000 input features (`params['model_arch_1']['dim_features'] = 2000`), 256 nodes per hidden layer (`params['model_arch_1']['dim_hidden'] = 256`), and three hidden layers (`params['model_arch_1']['n_hidden_layers'] = 3`), the corresponding decoder would then have three hidden layers, 256 nodes per hidden layer, and 2,000 output features. That is, both the encoder and decoders for a specific modality are always mirrored.

This is facilitated for both self-reconstruction and cross-modal reconstruction implementations. For example, if you set your first modality's encoder/decoder combination to 2,000 input and output features, even if the second modality encoder/decoder combination is set to 224 features, the cross-modal reconstruction would result in 2,000 features. This design enables the model to capture both modality-specific features and cross-modal relationships, essential for exploring correlations in multimodal data. Further, decoders allow for the realistic data imputation, or cross-reconstruction of high-dimensional feature space values. Moreover, the data reconstructions using a trained model have a denoising effect and can be used for the purposes of smoothing out data, which is particularly helpful with scRNA-seq data which is known to be noisy, sparse, and prone to dropout effects [65, 66].

Dual-Modality Framework with Multimodal Scalability

Although the current implementation focuses on dual modalities, the UniVI architecture is inherently scalable to K modalities, making it well-suited for future advancements in multimodal single-cell sequencing technologies. In our dual-modality setup, the model learns two latent spaces and generates four decoded outputs, corresponding to the following mappings:

- $x_1 \rightarrow z_1 \rightarrow \hat{x}_1$ and $x_1 \rightarrow z_1 \rightarrow \hat{x}_2$: Modality 1 reconstructing both itself and Modality 2.
- $x_2 \rightarrow z_2 \rightarrow \hat{x}_2$ and $x_2 \rightarrow z_2 \rightarrow \hat{x}_1$: Modality 2 reconstructing both itself and Modality 1.

Activation Functions

All hidden layers in UniVI employ the rectified linear unit (ReLU) activation function [67], defined in Equation 5. ReLU is chosen due to its computational efficiency and its ability to mitigate the vanishing gradient problem, which facilitates the training of deep neural networks. However, the activation functions in the decoder's final layers differ based on the reconstruction distributions used, as detailed in the Decoders section. These output functions are specifically tailored to match the statistical properties of each modality, ensuring appropriate reconstruction of the input data.

Loss Function

The UniVI loss function builds upon the standard evidence lower bound (ELBO) loss used in VAEs [33], adapting it for the MoE framework while incorporating additional terms to facilitate multimodal data integration. The baseline loss function, originally proposed by Shi et al. [35], is defined in Equation 6. The first term in this equation quantifies the log-likelihood of the reconstructed data, ensuring that the model accurately reconstructs the input while penalizing discrepancies. To further promote disentanglement of latent features, we introduce the β parameter [34], modifying the loss function as described in Equation 7. This adjustment enables greater control over the relative importance of the KL divergence term, which enforces alignment between the learned latent distributions and the prior. When using a standard normal prior $\mathcal{N}(0, I)$, this regularization encourages latent variables to occupy continuous and structured regions of the latent space, ultimately enhancing the model's ability to generalize and interpolate between data points.

Building upon this loss formulation, we introduce additional cross-modal KL divergence terms across all unique modality combinations, weighted by the γ parameter, as defined in Equation 8. This term serves as a tunable hyperparameter that, when increased beyond one, progressively emphasizes the cross-modal feature distribution KL divergence within the loss function. By adjusting γ, researchers can control the balance of objectives during training, allowing them to prioritize reconstruction accuracy, alignment of latent distributions with the prior, or consistency between the latent spaces of different modalities.

The comprehensive loss function of the UniVI model is defined in Equation 9. In this equation: K is the number of modalities. $q_{\phi_k}(z_k|x_k)$ represents the encoder, which approximates the variational posterior for modality k, parameterized by ϕ_k. $p_{\theta_{kj}}(x_j|z_k)$ is the decoder, responsible for reconstructing data from modality j based on the latent representation z_k, parameterized by θ_{kj}. $p(z_P)$ denotes the prior distribution over the latent variables. The term $\mathbb{E}_{q_{\phi_k}(z_k|x_k)}\left[\sum_{j=1}^{K} \log p_{\theta_{kj}}(x_j \mid z_k)\right]$ represents the expected log-likelihood (reconstruction loss), capturing how well the model can reconstruct all modality combinations from the latent space. The next term, $D_{KL}\big(q_{\phi_k}(z_k \mid x_k) \,\|\, p(z_P)\big)$, is the Kullback-Leibler (KL) divergence, which regularizes the latent space towards the prior distribution, controlled by the weighting factor β. Then the final term in the equation, $\sum_{\substack{j=1 \\ j\neq k}}^{K} D_{KL}\big(q_{\phi_k}(z_k \mid x_k) \,\|\, q_{\phi_j}(z_j \mid x_j)\big)$, enforces cross-modal latent space alignment by minimizing the KL divergence between the latent distributions of different modalities, weighted by γ.

In total, this loss function effectively balances reconstruction fidelity, latent space regularization, and cross-modal alignment, which are crucial for the integration of multimodal single-cell data. The derivation of this final loss function can be traced through Equations 7 to 9, showing the progression from the basic ELBO to the comprehensive UniVI formulation.

Data Partitioning

All data used for UniVI integration model training was automatically partitioned by the UniVI workflow into training, validation, and test sets with an 80%/10%/10% split, respectively. This can be edited by the user by specifying `params['train']['train_fraction']` to a float representing the desired training set data fraction (i.e., `0.80`) and `params['train']['val_fraction']` to a float representing the desired validation set data fraction (used for early-stopping etc.). The test set is assigned the leftover data after partitioning the training and validation sets.

Early Stopping

To prevent overfitting and enhance model generalization, UniVI utilizes early stopping [68] during model training. The training process monitors the validation loss and training is halted when no improvement is observed over a predefined number of epochs. For example, if the patience term is set as `params['early_stopping']['patience'] = 20`, before the model hits its predefined number of total epochs, `params['train']['n_epochs']`, if no improvement is seen during training over the course of any subsequent 20 epochs, the model stops training and keeps note of the best-performing model epoch.

This approach ensures that the model retains the parameters corresponding to the optimal validation performance. For the analyses in this manuscript, early stopping was set to `20` epochs. This can be changed by the user by specifying

`params['early_stopping']['patience']` prior to model training. This parameter takes in an integer reflecting the desired number of epochs before halting the training process if the validation loss does not improve.

Training Procedure

Model training was performed iteratively using backpropagation [69]. In each training epoch, the model processed the training data in mini-batches based on the specified value for the batch size hyperparameter. Following each epoch, the model's performance is evaluated on the validation set. The total epoch count can be altered by specifying the `params['train']['n_epochs']` parameter prior to model training, where the input of this parameter is an integer representing the total desired epochs for training. All the models trained in this manuscript used a value of `800` for this parameter and usually relied on the early stopping metric to halt training and avoid over—or under—fitting. Training ceased when the early stopping criterion was satisfied, preserving the model state with the best validation performance. Hyperparameters, including learning rate, batch size, number of latent dimensions (d_{latent}), and the β and γ parameters, were optimized through grid search and cross-validation to achieve optimal performance.

Evaluation Metrics

To systematically compare UniVI models across different parameter configurations, we primarily use the fraction of samples closer than the true match (FOSCTTM) metric [70], as defined in Equation 10. This metric quantifies the degree of manifold alignment between modalities in the learned latent space. Upon training completion, UniVI provides a built-in function for computing the FOSCTTM score. The process begins by invoking the `update_to_best_epoch()` function from the `UniVI.CrossMap` class, ensuring the evaluation is based on the best-performing model checkpoint. Subsequently, the function `calc_frac()` from `UniVI.external.evaluate_FOSCTTM` is used to compute the FOSCTTM score on the specific model's test set, thereby assessing manifold alignment performance on unseen data.

The FOSCTTM metric is computed using the Euclidean distance between sample embeddings in the latent space of each modality. Given that ground-truth sample correspondences across modalities are known, the Euclidean distances between each sample and its true counterpart in the opposing modality are recorded. The distances are then ranked, and the number of samples that are closer to a given sample than its true match are counted. Finally, the fraction of test samples with a closer match than the ground-truth counterpart is computed. This ensures an objective and interpretable evaluation of cross-modal alignment.

Beyond FOSCTTM, UniVI includes additional built-in functions to compute integration clustering performance metrics, allowing for a comprehensive evaluation of clustering quality in the integrated data space. These metrics capture distinct properties of the clustering structure and enable direct comparison of integration performance across different methods. The integrated space clustering performance is quantified using normalized mutual information (NMI) [71] (Equation 11), adjusted Rand index (ARI) [72] (Equation 12), and clustering accuracy (ACC) (Equation 13) (Figure 4). These metrics are calculated by implementing the `predict_celltype()` function imported from `UniVI.utilities._evaluate`, followed by the EvalMap() class function which is imported from `UniVI.utilities._evaluate` using a trained UniVI model object.

Together, these evaluation metrics establish a rigorous framework for assessing the performance of UniVI, both in comparing different model configurations and parameter choices within UniVI, as well as in benchmarking against other multimodal integration methods. This comprehensive evaluation ensures a robust assessment of both intra-model optimization and inter-method performance comparisons.

Methodology for CITE-Seq Experiments

We performed several experiments using UniVI models trained on CITE-seq data using Dataset 1 to demonstrate UniVI's capabilities for the purposes of batch correction, cell-cell alignment, denoising, and imputation. Furthermore, we utilized the testbed to evaluate UniVI's performance against existing state-of-the-art integration methods. The following subsections describe the methodologies employed in these experiments. Of note, these experiments are organized the same as they are in the Results section, with a separate overarching section for the experiments performed using CITE-seq data and the experiments performed using 10x Multiome data.

Data Integration

We performed a series of experiments with the overarching goal of showcasing UniVI's data integration capabilities. We first visualized both modalities of Dataset 1 prior to model integration in low-dimensional space via UMAP (Figure 2a, Supplemental Figure 8) and PCA (Supplemental Figure 8) using several Scanpy [54] functions. For the UMAPs, we used `scanpy.pp.neighbors()` prior to performing `scanpy.tl.umap()` to calculate the UMAP coordinates of the samples in Dataset 1. Following this, we plotted the data using `scanpy.pl.umap()` by passing `color = ['batch', 'ident']` to the function. Following this, to plot the PCA plots, we ran the `scanpy.tl.pca()` function to calculate the principal components (PCs) prior to `scanpy.pl.pca()` to plot them. We passed in the `color = ['batch', 'ident']` key to this function to color the PCA plots by modality and by cell type (Supplemental Figure 8). These initial visualizations provided a baseline for evaluating clustering and batch effects prior to integration.

After training the UniVI model with the selected features outlined in the Dataset 1 preprocessing methods, we integrated the data using the model's optimal epoch state post-model training. The integrated latent space was then visualized using UMAP and PCA with the Scanpy functions described above, with plots colored by modality and cell type as described. Enhanced mixing of modalities and improved separation of cell types in the latent space indicated effective batch correction and model training.

We additionally generated UMAPs as described previously of the full Dataset 1 latent representations using the Scanpy methods described above in the Batch Correction sections, highlighting cell type clustering at increasing levels of granularity (level1 with 8 cell types, level2 with 31 cell types and level3 with 58 cell types as assigned by Hao et al. [45]). This allowed us to visualize the granularity at which our method learns biologically-relevant clustering information and use as a reference for additional plots when looking at cell identities within specific clusters.

Cell-Cell Alignment in the Latent Space

We evaluated UniVI's capacity for aligning individual cells across modalities by leveraging joint measurements from Dataset 1 (Figure 2b). UMAP embeddings of the latent representations from the two modalities from the previous step are stored in the `UniVI.CrossMap` model class object's `adata1.obsm['X_umap']` and `adata2.obsm['X_umap']` variables, respectively. We created these plots using `_plot_umap_pair()`, which is built into the UniVI model object class function. We performed this three separate times using each `n` value from the $n_{\texttt{samples}} = [150, 100, 50]$. This plotting function samples paired cell measurements from both modalities iteratively from the input data and maps them each to the latent space. For each paired cell measurement, the sample number for each iteration is visualized in the plot. Additionally, this function draws dotted lines between each paired cell between modalities to keep track of cell-cell alignments in our model. We chose the `n` values for ease of visualization without the plot becoming too crowded with overlapped samples.

Next, we used another similar built-in UniVI trained model object class function called `plot_umap_pair()`. This was performed separately four times, once for each value in $n_{\texttt{samples}} = [200, 500, 1000, 2000]$ to visualize more paired samples in the latent space and allow for some overlap to look at general trends between samples. The plots are similar to the previously described plots, except for each paired sample, colored dots are plotted to the lower-dimensional latent space mapping. We reasoned that once you plot a high number of paired cells, it becomes difficult to see their associated sample numbers given that there is so much overlap. For both plots, we assigned red colored cells to scRNA data and blue colored cells to scADT data. These visualizations enable a qualitative assessment of UniVI's ability to align cells across modalities, ensuring that corresponding cells from different modalities are closely positioned in the latent space.

Denoising Using Learned Biologically Relevant Features

To investigate UniVI's denoising capabilities, we utilized the values from both modalities of Dataset 1 (Figure 3a). We first found all the overlapping features between the scRNA and scADT modalities which resulted in 23 total shared markers—CD93, CD83, CD68, CD40, CCR10, CD163, CD19, CD226, CD27, CD207, CD72, CD2, CD69, CD34, CD14, CD9, TIGIT, CD24, XCR1, CX3CR1, CD36, CLEC12A, and CD22. For each modality separately, we generated heatmaps using the `scanpy.pl.heatmap()` function and passed in the Dataset 1 input data as input. We passed in the list containing the markers above as `var_names`, then grouped cells in columns by their level1 cell identity using `groupby = 'ident'`. We next generated UMAPs as previously described using the Dataset 1 features prior to model application to compare them to both the reconstructions and the imputed data. All the UMAPs and heatmaps use the level1 cell identities with the same color scheme.

The trained UniVI model was then employed to reconstruct the data for each modality through both self-reconstruction and cross-modal reconstruction for all four reconstruction combinations using the model state from the best-performing epoch:

1. Reconstructed scRNA-seq features using scRNA-seq data.
   $x_1 \rightarrow z_1 \rightarrow \hat{x}_1$
2. Reconstructed scRNA-seq features using scADT data.
   $x_2 \rightarrow z_2 \rightarrow \hat{x}_1$
3. Reconstructed scADT features using scADT data.
   $x_2 \rightarrow z_2 \rightarrow \hat{x}_2$
4. Reconstructed scADT features using scRNA-seq data.
   $x_1 \rightarrow z_1 \rightarrow \hat{x}_2$

To highlight UniVI's ability to denoise data, heatmaps of the reconstructed shared data features were produced as previously described using each of the data reconstructions mentioned above. Further, using all the features from the reconstructions, we generated UMAPs as previously described and colored them with their corresponding level1 cell identities. We generated a similar plot, but using all 2,000 scRNA and 224 scADT features with the `UniVI.CrossMap.plot_recon()` class function, which can be called from a loaded-in trained UniVI model `CrossMap()` class object (Supplemental Figure 2a-c). This plot highlights a more global view of the denoising effect of our method, whereas the previous denoising plots generated largely focus on specific features.

Moreover, scatter plots comparing raw and denoised data were generated using the `UniVI.CrossMap.anal_corr()` class function, which includes a reference line ($y = x$) to illustrate the extent of data alteration and denoising effectiveness (Supplemental Figure 2d). Specifically, we calculated the per-cluster root mean squared error (RMSE) for all level3 cell identity clusters. We performed this calculation on the test set for trained model-reconstructed scRNA-seq data versus the unaltered scRNA-seq expression data (normalized and scaled).

Additionally, we looked at the latent space activations for the specified markers to tie the UniVI latent space to biological features. We used the latent space UMAP coordinates and colored them by expression intensity for each of the markers before denoising and after (Supplemental Figure 3). Three total UMAPs were generated for each marker—one colored with the original data expression intensity, one colored with the self reconstructed data's expression intensity, and one colored with the cross-reconstructed data's expression intensity. This was done for each modality separately, resulting in 69 total UMAP plots. These plots can be recreated using the UniVI CrossMap model object class function, `plot_features_compare()`, passing in `m = 'm1'` and `lst_features`—a list of all the features of interest (specified above). This will generate the described plots for the scRNA data since we ran the function with `m = 'm1'`. Next, to look at the latent activations for the scADT features, we can repeat the above steps, except we'll pass in `m = 'm2'` to specify modality. We can use the same `lst_features` in this case since we used overlapping features between modalities.

Further, to test the reconstruction accuracy against the original data, we calculated the RMSE of the reconstruction of each modality separately using the specific features mentioned above (Supplemental Figure 4). We first calculated the RMSE of all the reconstructed features versus the original features in Dataset 1 per individual cell. For each cell type in the level1 cell types, we plotted the mean of all the RMSE values separately and split them by the specific features used in this section. The resulting plot has each of the features on the X-axis, where each feature has a different bar plotted per cell type with representative colors. The Y-axis contains the mean of the RMSE for each specific feature for each specific cell type. We performed the plot generation outlined here twice, once for each modality and its corresponding self-reconstruction.

For a more holistic overview of the reconstruction accuracy provided by our method, we generated several supplemental plots. We first generated plots of the original and reconstructed count distributions for each modality separately (Supplemental

Figure 5). For each modality, we plotted the data distribution prior to reconstruction in blue and after reconstruction in green on the same coordinates. Moreover, we visualized the per-feature RMSE distributions for all the features used for model training in each modality separately (Supplemental Figure 6). Finally, to see the RMSE values across cell types in UMAP space to determine if high RMSEs were associated with specific cell types or populations of cells, we plotted UMAPs of the data modalities separately colored by RMSE intensity (Supplemental Figure 7).

Data Imputation

To show the ability of our method to impute realistic-looking, biologically relevant data, we used our trained model to impute data from both modalities separately (Figure 3b). We first used the scikit-learn [73] Python module and generated a Gaussian mixture model (GMM) of the original features used to train the model for each modality. Specifically for this step, we imported `GaussianMixture()` from `sklearn.mixture`. For each modality, we sampled from the GMM at random up to 100,000 samples—to ensure we had enough samples from each synonymous cell type for plotting—and then used the best-performing epoch of our UniVI model to reconstruct the input data. For this step, we performed scRNA to scRNA reconstruction and scADT to scADT reconstruction.

To infer cell type of the imputed samples, we used the reconstructed features and performed unsupervised clustering using `scanpy.tl.leiden()`, tweaking the `resolution` parameter until eight total clusters were generated to match the eight level1 cell types we started with. Next, we sampled 1,500 cells from each imputed "cell type" to match the Dataset 1 sample count for each level1 cell type. Moreover, we inferred cell types of the Leiden clusters using the previously derived marker expression patterns and gave each imputed sample an assigned level1 identity. Following this step, we used the same shared markers defined previously and generated heatmaps for each modality's imputed data. Further, we generated UMAPs as described previously with all the imputed features and colored them by their inferred cell type to visualize how the imputed samples clustered compared to the original data.

Integration Tool Evaluation

To benchmark UniVI against existing integration tools, we compared its performance using Dataset 1 with four widely-used methods: Seurat v3 (Anchor) [7], Seurat v4 (Weighted Nearest Neighbor (WNN)) [45], LIGER [57], and TotalVI [74] (Figure 4). Each method was applied as follows: For TotalVI, we used the default settings to train the model on the training set. The latent representations for the test set were extracted using the trained model via the `TotalVI.get_latent_representation()` function. For Seurat v3, we utilized `FindIntegrationAnchors()` and for Seurat v4, we utilized their synonymous function, `FindMultiModalNeighbors()` to generate their corresponding integrated representations. For LIGER, the `optimizeALS()` function was used, followed by `quantile_norm()` to create the integrated representations. Unless otherwise specified, all methods were run with author-recommended parameters.

Clustering of the integrated representations was performed using K-means clustering with `K = 8`, as well as `K = 31` corresponding to both the level1 and level2 cell type class count, respectively. We implemented K-means clustering using the `sklearn.cluster.KMeans()` function across the 20 replicates to ensure robust cluster assignment for comparison. We generated UMAP visualizations of the integrated layers using the `scanpy.pp.neighbors()` function, followed by `scanpy.tl.umap()`. We colored these visualizations by K-means cluster assignments, batch, donor, predicted clusters, level1 cell identities, and level2 cell identities. We generated the plots using the `scanpy.pl.umap()` function with the color argument specifying the desired annotation.

To interpret clustering performance, we constructed Sankey plots to visualize the correspondence between predicted K-means clusters and true annotated level1 cell identities for the test set. We generated these plots using the `plotly.graph_objects.Sankey()` function, where node labels represented K-means clusters and level1 cell identities.

Furthermore, we performed quantitative evaluation by computing NMI, ARI, and ACC of predicted K-means clusters versus ground truth level1 cell identities. We calculated these metrics using the following `sklearn.metrics` functions, `normalized_mutual_info_score()` for NMI, `adjusted_rand_score()` for ARI, and a custom function for ACC using the `unsupervised_clustering_accuracy()` function, imported from `UniVI.utilities._evaluate`. These functions are all performed using a UniVI wrapper function by first running the `utilities._evaluate.predict_celltype()` function, followed by the `utilities._evaluate.EvalMap()` class function. We computed each metric for all 20 independent replicates of Dataset 1, and the results were averaged and visualized using error bars in box plots.

Methodology for 10x Multiome Experiments

We next conducted experiments using UniVI models trained on the 10x Genomics Multiome datasets: Dataset 2 and Dataset 3. The analyses were categorized based on the dataset used for training, employing specific methodologies for each dataset type. The following subsections describe the approaches utilized in these experiments. We also used Dataset 4 and Dataset 5 in the experiments specified below using a UniVI model trained on Dataset 3.

Deep Learning Classification Using UniVI Cross Modally-Generated ATAC Data

Utilizing Dataset 2, we assessed UniVI's ability to both integrate 10x Multiome data and generate scATAC-seq features from scRNA-seq data and tested how transferrable the generated data was across modalities (Figure 5). A UniVI model was first trained using the data split set aside for model training/validation/testing. We plotted UMAPs of the integrated training data as previously described (Supplemental Figure 8). The Dataset 2 20% split was integrated using the pre-trained model after filtering to include only the features used during training.

We generated UMAP visualizations of the models' latent space, colored by modality and cell type using techniques previously described (Supplemental Figure 8). We also used the Dataset 2 test split and integrated the data, visualizing the latent space using UMAPs colored by modality and cell identity. The pre-trained model was then used to generate scATAC-seq data feature reconstructions from the test set scRNA-seq data. To visually inspect the similarity of the cross-modally reconstructed data and the actual data, we generated UMAPS of the scATAC-seq features along with the real scATAC-seq data from the test set. We colored one UMAP by modality using the 'scRNA' label to denote the scATAC-seq data reconstructed from the scRNA-seq data and the 'scATAC' label to denote

the actual scATAC-seq test data. Additionally, we visualized the same UMAP embedding, except colored by cell type to determine if cell types were clustered together between the datasets.

We employed the Dataset 2 20% split's scATAC-seq data and the cross-generated scATAC-seq data to train and evaluate deep learning classification model. This analysis was aimed at predicting cell types from the unaltered scATAC-seq data using a model trained on cross-modally generated data. To better learn classification representations for minority classes, we oversampled up to 50 using a random selection process for both modalities together if a class contained less than 50 samples. We utilized PyTorch [63] for two different classification tasks as follows:

1. **High-Resolution Cell Type Classification** Which involved 19 cell type classes. The neural network architecture comprised an input layer matching the number of generated scATAC-seq features (5,000), followed by two fully-connected hidden layers with 516 and 256 neurons, respectively. Each hidden layer was followed by batch normalization, ReLU activation, and 50% dropout. The output layer corresponded to the 19 cell type classes. The network was trained using the Adam optimizer with a learning rate of 0.001 over 30 epochs.
2. **Low-Resolution Cell Type Classification** Which involved seven cell type classes. A similar neural network architecture was employed, with the output layer adjusted to match the reduced number of classes.

Post-training with the cross-generated data, we tested the models using the actual scATAC-seq peak data. We calculated and visualized performance through confusion matrices, total accuracy metrics, and Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) [75] plots. We created custom code to generate these plots using both the Matplotlib [76] and Seaborn [77] Python packages. To ensure statistical robustness, we performed 10-fold cross-validation for each classification experiment and average accuracies were reported across folds.

Integrating Disparate Unimodal Datasets Using a UniVI Model Trained on Multimodal Data

We evaluated UniVI's capability to integrate separately measured and processed unimodal scRNA-seq data (Dataset 4) and scATAC-seq data (Dataset 5) using a model trained on Dataset 3 (Figure 6). Once trained, we loaded the `UniVI.CrossMap()` class object and updated the model to the best-performing epoch. This was done using the built-in class function, `update_to_best_epoch()`. After we updated the model to this epoch, we used the `UniVI.CrossMap()` object's inherited class function `get_latent_features()` to generate the latent representation of both Dataset 4 and Dataset 5 separately. To visualize Dataset 4 and 5 prior to integration, we separately generated UMAP plots for each unimodal data modality, first colored by technology, then by cell type for each modality.

To make the features more comparable between datasets for the integration visualizations, we mapped cell types to matching categories between unimodal datasets. Some of the cell types between datasets did not have a cross-modality counterpart, such as 'Regulatory T' cells in the scATAC-seq modality and 'Cytotoxic T' cells in the scRNA-seq modality. Therefore, a few of the annotations only have representative samples from a single modality. For visualization purposes, we removed cells with ambiguous or uninformative annotations, such as 'Unassigned', 'PBMC', and 'Bone Marrow' since these are relatively non-specific and were cluttering our combined visualizations. This resulted in a total of 19,567 scATAC-seq cells and 30,449 scRNA-seq cells. Next, we generated UMAPs using the integrated latent representations of both datasets colored by 'tech' and 'celltype'.

To quantify the distances between modalities for similar-cell types, we identified cluster centroids in the latent manifold by cell type for both modalities. We created a heatmap displaying the Euclidean distances between cluster centroids for each modality across all 60 latent dimensions, with a darker blue color indicating closer proximity. This visualization provided quantitative insights into the integration effectiveness, highlighting the clustering of similar cell types across different unimodal modalities using a model trained on 10x Multiome data.

Reconstructing scRNA-Seq Data to Correct Dropout Effects

To assess UniVI's denoising capabilities in the context of scRNA-seq data dropout, we used a method similar to the work described in the scVI manuscript [16] (Figure 7). For this experiment, we split 80% of the data as described to be used for the UniVI model and retained 20% of the data to perform the denoising experiment discussed here. We first simulated dropout in the scRNA-seq data of the test set by randomly setting a proportion of non-zero counts to zero across various percentages (10% to 95%). The dropout simulation was performed across the following proportions: `percentage_list = [.10, .20, .30, .40, .50, .60, .70, .80, .90, .95]`. The data was scaled to Z-scores to align with the model's input requirements.

We evaluated reconstruction performance by using the trained UniVI model to reconstruct the unaltered test set, the test set with 20% simulated dropout, and the test set with 50% simulated dropout. We generated heatmaps for the top 1,000 variable mRNA features out of the available 2,000 for clarity, comparing input data with reconstructed data across different dropout levels. To quantify reconstruction accuracy, we calculated the absolute differences between reconstructed and original data for the dropped-out features. Mean and median reconstruction errors were plotted against the corresponding dropout percentages to illustrate the model's denoising effectiveness across varying levels of data loss.

Furthermore, to evaluate the preservation of biological signals, we performed Gene Set Enrichment Analysis (GSEA) using GSEApy [78] on ranked gene expression features from the intact, dropout-simulated, and reconstructed datasets. Enrichment scores for KEGG pathways [79] and Gene Ontology (GO) [80] Biological Processes (BP) were calculated. We computed Pearson correlation coefficients between the enrichment scores of the intact data and those of the dropout-simulated and reconstructed data. We visualized these correlations using bar graphs to demonstrate the extent to which biological pathways were preserved following denoising and imputation.

Results

In this section, we highlight the results from all experiments performed in this manuscript. Of note, the experiments are organized by the overarching multimodal data type utilized.

The first section, "UniVI Can Align Joint Multimodal Datasets Without Reliance on Identity Information Across Modalities" and all its corresponding subsections contain results from experiments that were performed using models trained on CITE-seq data (Dataset 1). While the second section, "Expanding UniVI's Reach: Application to 10x Genomics Multiome Data" and its subsections contain results from experiments that were performed using models trained on 10x Multiome data (Dataset 2 and Dataset 3).

UniVI Can Align Joint Multimodal Datasets Without Reliance on Identity Information Across Modalities

We evaluated UniVI's performance by treating joint multimodal datasets with known alignments as disjoint and assessed the cross-modal reconstruction error and proximity of reconstructed instances to their counterpart in the opposite modality. For this analysis, we used Dataset 1, a robust CITE-seq dataset (Materials and Methods).

We first confirmed that UniVI maps the two modalities to overlapping latent spaces (**Figure 2a**). For initial visualization of the unintegrated data, we generated UMAP (Figure 2a, Supplemental Figure 1) and PCA plots (Supplemental Figure 1) on normalized and scaled CITE-seq data post-feature selection. We generated plots to highlight both modality and level1 cell identities. Next, we trained UniVI on a large subset of the data and integrated both modalities with one-another. Using the best-performing epoch of the trained model, we calculated the mean from each node of the cross-modal latent space of all data samples, and visualized these representations with UMAP for modality, level1 cell identity, level2 cell identity, and level3 cell identity.

Post-integration, UniVI effectively eliminated modality-specific batch effects while clustering cells by their biological identities without using prior information. The integrations highlight the ability of our method to use a relatively small subset of data to learn the important features for clustering by the most relevant sources of variation—biological cell identity. We extended our analysis to evaluate latent alignment performance at finer cell type granularities to highlight the integration resolution. UMAPs of cross-modal latent representations were generated with detailed cell type annotations, defined as level2 and level3 identities by Hao et al. and related studies [7, 8, 45, 81]. While level1 identities represent general simplified cell types, level2 and level3 provide higher-resolution annotations, with 31 and 58 cell types, respectively.

The results indicate that most high-resolution cell types cluster accurately in the latent space, reflecting refined biological differences. This enables cross-modal visualization of biological variation and cell lineage relationships, as spatial proximity in the latent space reveals related or merging populations. These findings demonstrate that UniVI effectively captures highly detailed cross-modal biological information, even at the most granular cell type annotations. The latent space shows that our model captures even the most complex sources of biological variation using a limited feature set.

Precise Cell-Cell Alignment Demonstrates UniVI's Accuracy

To further interrogate our method's ability to accurately cluster cells in the latent space across modalities, we analyzed cross-modal alignment between identical cells (**Figure 2b**). Using the previously trained model, we used the means of the cross-modal latent representations of the Dataset 1 input vectors. Since the dataset includes jointly measured cells, the ground truth enables assessment of cross-modality manifold alignment at single-cell resolution.

To visualize integration and paired sample correspondence, we mapped identical cells with dotted lines on UMAP projections for various limited sampling levels. The first plot aligns individual cells and numbers them by their specific number in the sampling order used for this analysis. For example, cell 29 in each color are different modality measurements from the same cell which we connected with a dotted line. It was the 29th cell sampled from our data for this visualization. 150 samples were chosen here to limit cluttering within the figure. We then made two less complex versions of this plot and "peel away" samples, leaving a second plot with 100 samples, and a third plot with only 50 samples for visual clarity.

It is important to note that in UMAP space, the relative distances between points should not be over-interpreted as a precise measure of similarity. Instead, UMAP is most informative in terms of clustering, where local neighborhood structure reflects shared biological variation among cells. In the context of our multimodal alignment, the focus should be on how well similar cells cluster together rather than the perceived length of the distances between them. With this in mind, as long as cells align within the same biological cluster, that should largely be considered a success for single-cell data integration methods when using this technique. We see that is largely the case for all samples visualized. We do see some examples in which the two modalities do not seem to align for the same cell. For instance, cell 59's scRNA-seq data seems to cluster closer to a Dendritic Cell than its scADT counterpart measurement. Moreover, it clusters closer to the B cell cluster, illustrating a case where the alignment between the two modalities is inconsistent. Although, we cannot be certain if this phenomenon stems from actual biological variation, batch effect, integration error, or from some other cause.

We next generated four different UMAP plots with colored points that represent single-cell samples for `n_samples = [200, 500, 1000, 2000]`—where `n_samples[i]` is the total cell count for each visualization—in order to view more general trends between modalities. As we increase in sample size, our aligned latent manifolds do show some cells that are clustered across cell types. But this phenomenon is relatively rare, appearing to occur in roughly 1% of samples when looking at the plot with the highest sample count (`n_samples[i] = 2000`).

This raises the consideration that cellular states are inherently dynamic and continuously evolving, making it plausible that the integrated latent spaces capture genuine biological signal rather than errors in alignment—although a definitive answer to this would require further analysis. Additionally, observed discrepancies may stem from technical artifacts inherent to sequencing technologies rather than limitations in the integration approach. Notably, our model achieved a FOSCTTM score of 0.104 for this experiment, indicating strong overall alignment performance. These visualizations further confirm that most corresponding cells are mapped correctly, particularly among phenotypically similar cell types. Furthermore, alignment is consistently preserved within well-defined identity cluster regions, reinforcing the reliability and robustness of our multimodal data integration framework in capturing shared biological structure across modalities.

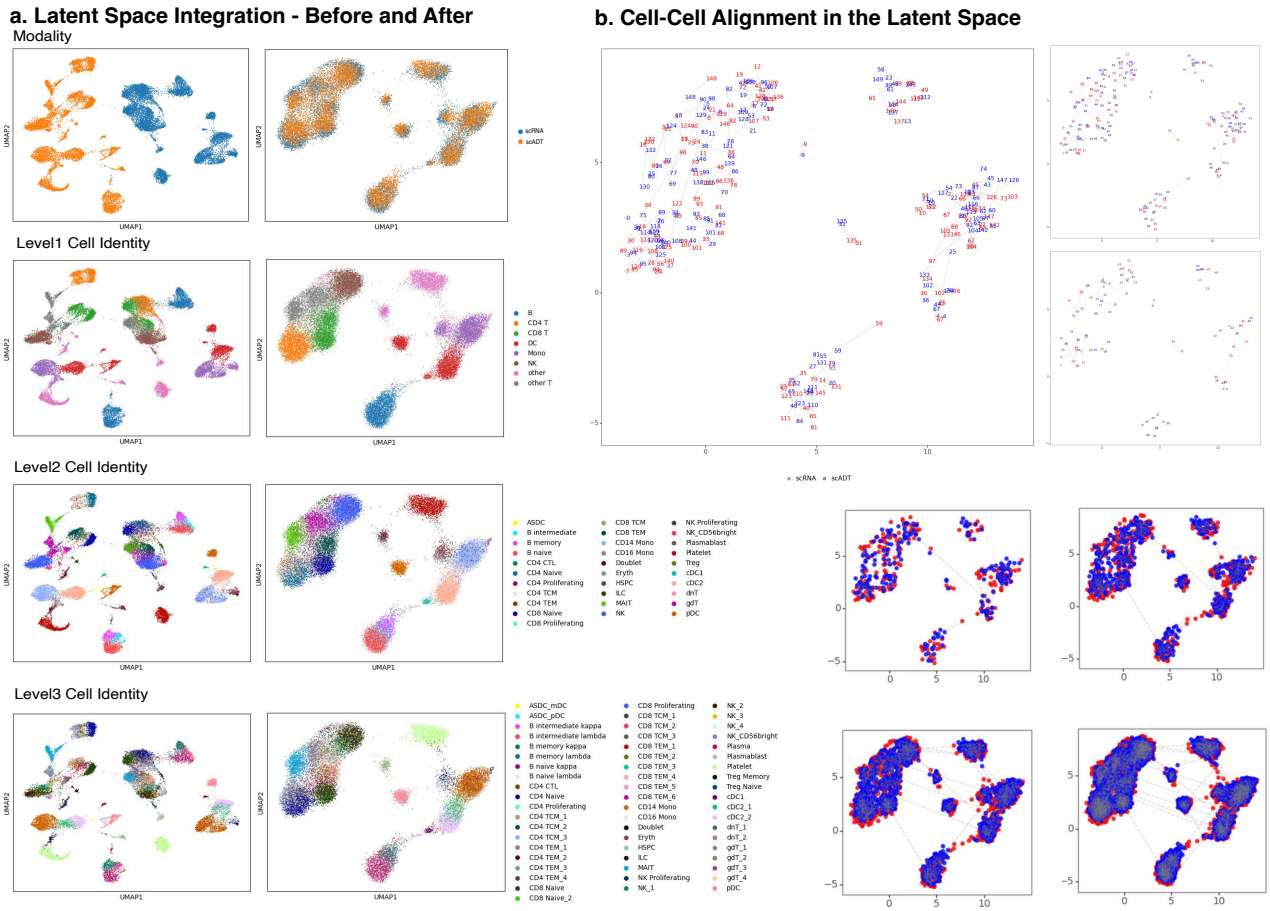

Fig. 2. Integration Results Using Single-Cell CITE-Seq Human PBMC Data. We highlight the integration performance of UniVI when unifying CITE-seq data. The UMAP visualizations here include samples mapped to the trained models' latent spaces. **(a.)** (*Top*) We show UMAP plots of both modalities before and after integration, we show that the batch effects between data modalities are removed using our method and all samples are clustered due to other sources of variation. (*Second Row*) Next, we colored the same UMAP plot using the low-granularity cell type annotations. We show that both modalities are overlapped and clustering by cell type. (*Third Row*) Next, we highlight the data prior to—and after—integration, colored by higher-granularity cell type annotations. We show here that the integrated data clusters continue to show separation, underscoring our models ability to learn complex, biologically relevant features. (*Bottom*) We further spotlight that the clusters in our integrated data space between modalities shows continued separation at the highest-granularity cell identity annotation, accentuating our method's ability to integrate complex data. **(b.)** We next underscore alignment between modalities at the single-cell resolution. (*Top*) We feature a small portion of the data for visual clarity (n = 150, 100, then 50, top to bottom, left to right), numbered by cell sample number in the integrated data space and exhibit dotted-lines between joint samples from both modalities. This showcases our model's manifold alignment at the single-cell resolution. (*Bottom*) We continue to show joint samples with dotted-lines mapped between CITE-seq modality measurements, plotting dots instead of cell sample number for the remaining figures in the panel. Moreover, we increase n for each plot from left to right, top to bottom, with n = 200, 500, 1,000, then 2,000. Taken together, we show that UniVI accurately integrates multimodal CITE-seq data, clustering both modalities up to the highest-granularity cell identities. We further underscore that at the individual cell level, our method maps joint samples with great accuracy between learned manifolds.

Enhancing Data Clarity: UniVI's Effective Denoising Capabilities

We observed that UniVI exhibits strong denoising capabilities after data reconstruction, effectively smoothing data by refining biologically relevant signals and dampening noise. This property is particularly valuable in the context of scRNA-seq datasets, where dropout events and undersampling artifacts are common [66, 82]. While autoencoders have been widely applied for unimodal denoising tasks [83], the introduction of cross-modal constraints within UniVI further improves noise reduction and preserves biologically meaningful variation across data modalities [84].

To assess UniVI's ability to recover biologically meaningful signals, we focused on all shared features between the scRNA-seq and scADT modalities. This allowed us to not only make comparisons within modalities, but across modalities as well. Moreover, since the scADT features were largely chosen as cell identity-defining markers, several of these features allow for the inference of cell identities in a similar manner for both modalities. Heatmaps of reconstructed data reveal that UniVI data reconstructions accurately and effectively smooth expression profiles, reducing sparsity while maintaining strong cell type specificity (**Figure 3a**, Supplemental Figure 2). For instance, markers highly expressed in monocytes, such as CD14 and CD68 [85, 86], remained distinct post-reconstruction, while their expression appeared clearer and less fragmented. Likewise, CD19, a well-characterized B cell marker [87], exhibited improved signal consistency in the B cell cluster across both modalities following denoising.

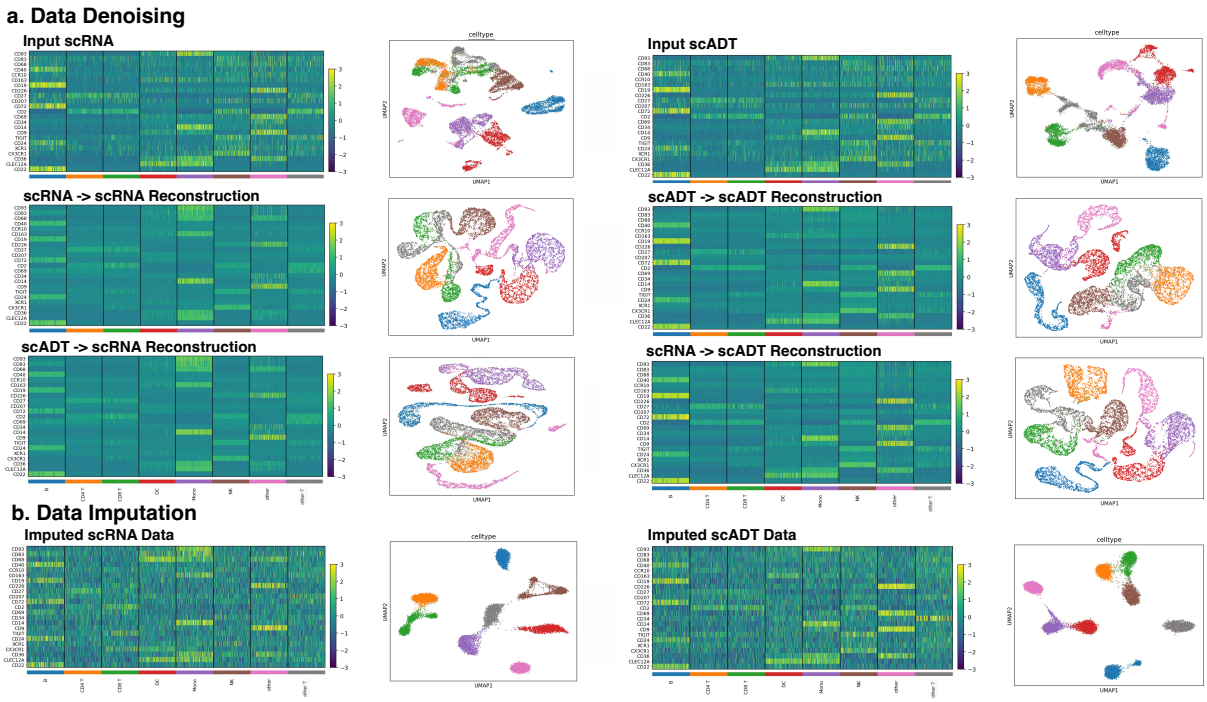

Fig. 3. Biologically Relevant Denoising and Imputation Using Learned Features. We spotlight that we can use the learned features to effectively denoise data unimodally and cross-modally, visualizing all the shared features between CITE-seq modalities. Moreover, we highlight that we can impute realistic-looking samples for either data type. As the colors for the heatmaps increase from hues of blue to green to yellow, it indicates increasing expression intensity from low to high. Moreover, we included the UMAPs of the data represented with the same color scheme as the heatmap cell types. **(a.)** Heatmaps of overlapping features across cell types and modalities. Top-left panels from top to bottom compare raw and UniVI-reconstructed expression. (*Top Left*) We first show the scRNA-seq shared features prior to model reconstruction. (*Middle Left*) Next, we highlight the scRNA-seq data reconstructed by the model using the scRNA-seq data. We see that this smooths out the learned features, which is especially significant for this modality since it is known to be sparse and prone to dropout effects. (*Bottom Left*) We next highlight the denoising capability of our model via cross-generation of data—that is—we used the scADT data to generate scRNA-seq features. These plots highlight that UniVI-reconstructed data presents a smoothing effect that maintains higher expression of important cell lineage-defining features. Furthermore, we highlight similar plots previously described, except here looking at the shared features in the scADT data modality. (*Top Right*) We visualize the scADT protein expression prior to feature reconstruction using our method. (*Middle Right*) We feature the scADT proteins reconstructed, which manifests a similar smoothing/denoising effect for this modality. (*Bottom Right*) Lastly, we emphasize the denoising effect that occurs when we cross-modally generate data, in this case using scRNA-seq transcript expression to reconstruct the scADT protein levels. All highly expressed markers are confirmed in the literature to be associated with their corresponding cell identities. **(b.)** We next highlight heatmap plots that contain the same shared features, except the samples highlighted in this panel for each modality are synthetic. We additionally show UMAPs of the imputed features to show how the data are clustered. To generate these samples, we randomly sampled from the original feature value distributions and used a UniVI model to reconstruct the features. (*Left*) We show that the synthetic scRNA-seq data displays increased expression of cell type-defining markers with added feature variation and noise that resembles the input data prior to model application. Moreover, the imputed samples cluster together similar to the input data by their corresponding "cell type". (*Right*) We showed a similar phenomenon with the synthetic scADT samples we generated, with cells clustering together that we would expect given the input data. Additionally, there is variational noise in the imputed samples, which we also see in the input data. Overall, UniVI not only unifies multimodal CITE-seq data but also leverages the learned latent features for biologically meaningful denoising and realistic sample imputation. By smoothing out noise in both unimodal and cross-modal reconstructions and accurately generating synthetic data that preserves key cell type marker patterns, UniVI demonstrates its power to refine single-cell measurements and deepen biological insights.

Interestingly, UniVI's cross-modal reconstructions not only preserve key lineage markers but also enhance the separation of closely related cell populations in a biologically meaningful way. For example, CD27, which is known to distinguish subsets of B cells, including memory B cells [88], was reconstructed with increased clarity in a subset of B cells while remaining less prominent in others (Supplemental Figure 3). This pattern suggests that UniVI is capturing biologically relevant heterogeneity within the B cell population. Similarly, CD83, which is associated with dendritic cells and monocytes [89], was retained in these clusters post-reconstruction, with improved separation from other immune lineages. These are just a few examples of this phenomenon, but we see this pattern across several of the markers visualized.

To quantitatively evaluate UniVI's denoising performance, we computed the RMSE for each marker within each cell type separately (Supplemental Figure 4). This analysis was designed not only to assess reconstruction accuracy but also to contextualize error distributions across distinct features and cell types. Notably, reconstruction error varied across both features and modalities, with scRNA-seq reconstructions exhibiting higher overall RMSE values (maximum ≈ 2.25) compared to scADT reconstructions (maximum ≈ 1.7). This disparity likely reflects the greater sparsity and pronounced dropout effects characteristic of transcriptomic data.

Moreover, markers with higher expression levels in a given cell type tended to exhibit greater reconstruction errors, likely due to the substantial dynamic range of these features and the pronounced dropout effects in scRNA-seq. In contrast, markers

with consistently low expression across cell types showed lower reconstruction errors, as differences between reconstructed and observed values were less pronounced. These trends were consistent across both modalities, reinforcing UniVI's ability to model biologically relevant expression patterns while underscoring the challenges associated with recovering highly variable transcriptomic features in sparse single-cell data.

To further assess the biological coherence of denoised data, we examined reconstructed feature expression at three stages: pre-integration, self-reconstruction, and cross-reconstruction. Heatmaps of key lineage-defining markers, such as CD14 (monocytes) [85], CD19 (B cells) [87], and CD2 (T and NK cells) [90], demonstrated that these features remained well-preserved post-reconstruction. Notably, UniVI reduced noise while maintaining biologically relevant variation across immune lineages. For instance, CD69—a known marker of activated T cells and NK cells [91]—exhibited a sharper distinction in reconstructed data, further highlighting UniVI's ability to enhance resolution within marker-specific gene expression profiles.

Examining the UMAP embeddings in conjunction with the heatmaps, we observed that UniVI maintained robust cluster separations across immune cell types, even after integration and denoising. Importantly, reconstructed data exhibited greater delineation between related subsets, particularly within the CD69-expressing population. These results suggest that UniVI not only preserves biological integrity but also enhances the interpretability of single-cell multimodal data by reinforcing meaningful structure within integrated latent spaces.

To further evaluate the impact of UniVI's denoising effect, we first examined the distribution of feature counts before and after reconstruction (Supplemental Figure 5). The post-reconstruction distributions revealed a notable smoothing effect, particularly in scRNA-seq data, where dropout effects are prevalent. Prior to reconstruction, the feature count distributions exhibited a sharp drop-off near zero, indicative of a high proportion of undetected values—potentially due to technical limitations. After denoising, we observed a more continuous spread of feature values, with a transition from near-zero counts to low-intensity values. This shift suggests that UniVI effectively mitigates dropout effects, better preserving biological signal across cell types and improving feature resolution. A similar, though less pronounced, effect was observed in scADT data, where protein expression values also exhibited a smoother distribution following reconstruction.

To quantify the fidelity of feature reconstruction, we computed the RMSE across all features (Supplemental Figure 6) and identified the 100 features with the highest and lowest RMSE values (Supplemental Table 1). Features with the highest RMSE predominantly belonged to ambiguous or rare cell types, where biological variation is inherently greater, presenting a challenge for precise reconstruction (Supplemental Figure 7). This suggests that UniVI's reconstruction error is not uniformly distributed across features but instead reflects biologically meaningful differences in expression dynamics. Collectively, these results demonstrate UniVI's capacity to denoise single-cell data while preserving biologically relevant marker expression, facilitating clearer cell-type distinctions and improving multimodal data interpretability for downstream analyses.

Leveraging UniVI for Synthetic Single-Cell Data Generation and Imputation

Given that UniVI is fundamentally a generative model, we sought to assess its ability to produce biologically plausible synthetic single-cell data samples. While UniVI inherently supports multiple modes of data imputation—including cross-modal sample generation, where an input from one modality is projected into the shared latent space and subsequently decoded into its cross-modal counterpart—we specifically aimed to assess its capability to generate synthetic samples for each modality separately de novo. By leveraging the aligned latent manifold, UniVI enables structured interpolation within the biological feature space, making it a valuable tool for generating realistic single-cell expression profiles. To implement this, we combined our method with a GMM to facilitate probabilistic data imputation (**Figure 3b**).

To accomplish this, we first trained a GMM on the input feature space, effectively modeling a simplified approximation of the structure present in the original dataset for each modality. While the GMM captures broad feature distributions, it lacks the fine-grained biological complexity inherent to real single-cell data. To introduce this complexity, we leveraged our method's deep generative modeling capabilities. The synthetic feature vectors sampled from the GMM were first encoded into the UniVI latent space, where the model's learned manifold imposed biologically meaningful structure onto the data. Within this space, additional stochastic sampling was performed, allowing UniVI to introduce nonlinearity and capture the intrinsic variability observed in real single-cell measurements—even adding realistic "noise" to the data. Finally, the enriched latent representations were decoded back into the high-dimensional expression space, where UniVI's decoders reconstructed feature distributions with realistic variation across cells. This process transforms the initially coarse GMM-derived feature vectors into biologically coherent, high-fidelity synthetic samples that retain key properties of the original dataset. By refining and enriching the simplified synthetic data through latent space transformations, UniVI enhances the plausibility of generated samples, demonstrating its utility for biologically informed data augmentation, imputation, and de novo single-cell whole sample generation.

To assess the biological fidelity of the imputed samples, we compared their feature expression intensity to those of real cells using heatmaps of all the shared markers between modalities for both continuity and comparison. The imputed data exhibited realistic marker expression patterns that aligned with known immune cell identities. For instance, key B cell markers such as CD19 and CD22 were appropriately restricted to B cell-like populations, whereas monocyte-associated markers such as CD14 and CD163 remained confined to myeloid-like clusters [85, 92]. Similarly, T cell markers including CD27 and CD69 were predominantly enriched within the imputed T cell populations, consistent with their expected biological specificity [88, 91].

To further evaluate the biological relevance of imputed samples, we concurrently projected them into UMAP embeddings alongside both the original and reconstructed data from the previous experiment. The imputed samples consistently localized within biologically meaningful regions of the UMAP space, clustering with their expected cell populations rather than forming spurious or ambiguous groupings. Notably, cell types with greater phenotypic similarity exhibited closer proximity in the embedding,

reflecting preserved biological relationships. This indicates that UniVI's learned latent manifold effectively captures and retains critical variation underlying cellular identity, ensuring that imputed samples align with real biological structures and remain interpretable within the broader dataset.

We've shown here that we can impute biologically plausible whole-sample data using our method in conjunction with a GMM modeled after the input data. At a more granular—and perhaps more practically useful—level, we can leverage this capability to impute just those features that are missing for a particular dataset. This could address technical limitations such as dropout effects in scRNA-seq [66, 82]—effectively rescuing attenuated transcript expression—or sparsely measured protein abundances in CITE-seq scADT data—which often suffers from background noise, limited dynamic range, and batch-dependent variability [41, 93].

By providing a partially observed sample as input, UniVI can infer the corresponding missing values by mapping the sample into its learned latent space and reconstructing the full feature set through either unimodal, cross-modal, or multimodal decoding. This enables the prediction of unmeasured transcriptomic or protein expression levels based on shared structural relationships across modalities, effectively filling in missing data while preserving biologically meaningful variation. Such an approach enhances the interpretability of multimodal single-cell datasets and expands the analytical potential of studies where comprehensive measurements across modalities are not always available. Overall, the ability to generate high-fidelity synthetic data at both the feature and whole-sample level has significant implications for single-cell research, particularly in cases where technical dropout and data sparsity limit downstream analyses.

Comparative Evaluation of UniVI's Multimodal Integration Performance

To systematically evaluate UniVI's integration performance, we benchmarked it against four widely used multimodal integration methods: Seurat v3 (Anchor) [7], Seurat v4 (WNN) [45], LIGER [57], and TotalVI [74] (**Figure 4**). Each method was applied to Dataset 1 using author-recommended parameters (Materials and Methods). Given the absence of a universally accepted single metric for multimodal data integration, we employed a suite of complementary evaluation criteria, each capturing different aspects of cluster coherence and biological fidelity. Specifically, we quantified integration performance using NMI [71], which measures how well predicted clusters preserve known cell annotations, ARI [72], which evaluates clustering accuracy by comparing predicted groupings to ground-truth labels, and ACC, which quantifies direct alignment between predicted and true labels using the integrated data space. Specifically, this metric measures how well the unsupervised K-means clustering method can predict cell type given the integrated data space, then ACC is calculated from the K-means predicted samples and the ground-truth labels. To ensure statistical robustness, each metric was computed over 20 independent validation experiments, maintaining consistent proportional representation of cell types across subsets (Materials and Methods).

To visually assess integration quality, we projected the learned embeddings into UMAP space, coloring them by batch, donor identity, predicted K-means clusters, and level1/level2 ground-truth annotations. Additionally, Sankey plots provided a detailed comparison between predicted and actual cell identities, revealing each method's ability to preserve biologically meaningful relationships and minimize erroneous label mixing. By leveraging these visualization approaches, we aimed to capture both global integration trends and more nuanced cell-type-specific alignment differences. This approach allowed us to examine whether UniVI maintained biologically relevant structures while mitigating technical artifacts that could arise from batch effects or inconsistencies in feature space alignment. Evaluating these visual representations alongside quantitative metrics provided a comprehensive view of how each method performed under real-world conditions.

Among the evaluated methods, UniVI achieved the second-highest NMI score, marginally trailing TotalVI. This result aligns with expectations, as TotalVI explicitly models protein and RNA data with modality-specific priors, giving it an advantage in CITE-seq data integration. However, UniVI surpassed all other methods—including TotalVI—in ARI and ACC, demonstrating its superior capacity to maintain fine-grained cellular distinctions while achieving seamless cross-modal alignment. Notably, while NMI provides a measure of overall cluster similarity, ARI and ACC more directly assess the accuracy of individual cell assignments, reinforcing UniVI's ability to maintain detailed and biologically relevant subgroup structures. The higher ARI and ACC scores further suggest that UniVI preserves within-cluster homogeneity more effectively than alternative methods, reducing the likelihood of spurious mixing or excessive smoothing that could obscure meaningful biological variation.

Further analysis of label transfer performance emphasized UniVI's robustness in aligning predicted K-means clusters with true annotations. Sankey plots further revealed that UniVI maintained higher consistency between predicted and ground-truth identities compared to Seurat-based methods and LIGER, both of which exhibited greater cluster fragmentation. These results indicate that UniVI effectively balances integration and biological fidelity, aligning related cell types while preserving meaningful heterogeneity within populations. The ability to correctly associate reconstructed cell types with their original annotations is particularly critical for downstream biological interpretation, as improper alignment can confound analyses and lead to inaccurate conclusions about cell state transitions and differentiation trajectories. By maintaining cell identity consistency while still allowing flexibility in latent space representation, UniVI ensures that biologically meaningful relationships remain intact throughout the integration process.

Unlike methods that impose strong modality-specific assumptions, such as TotalVI's generative protein modeling or Seurat's predefined anchor-based correspondence matching, UniVI operates without prior knowledge. This adaptability is particularly advantageous for datasets beyond CITE-seq, where predefined correspondences may be absent or unreliable. Furthermore, its architecture allows for robust generalization to new multimodal datasets, enabling researchers to apply the model across diverse experimental conditions without requiring extensive parameter tuning. As multimodal single-cell sequencing technologies continue to evolve, methods that maintain flexibility across different modalities and experimental setups will be increasingly valuable, making UniVI a promising candidate for broader applications in the field.

Collectively, our findings highlight UniVI's competitive or superior performance relative to established integration methods, underscoring its robustness and flexibility across diverse multimodal sequencing datasets. By providing accurate cell type annotation, biologically meaningful batch correction,

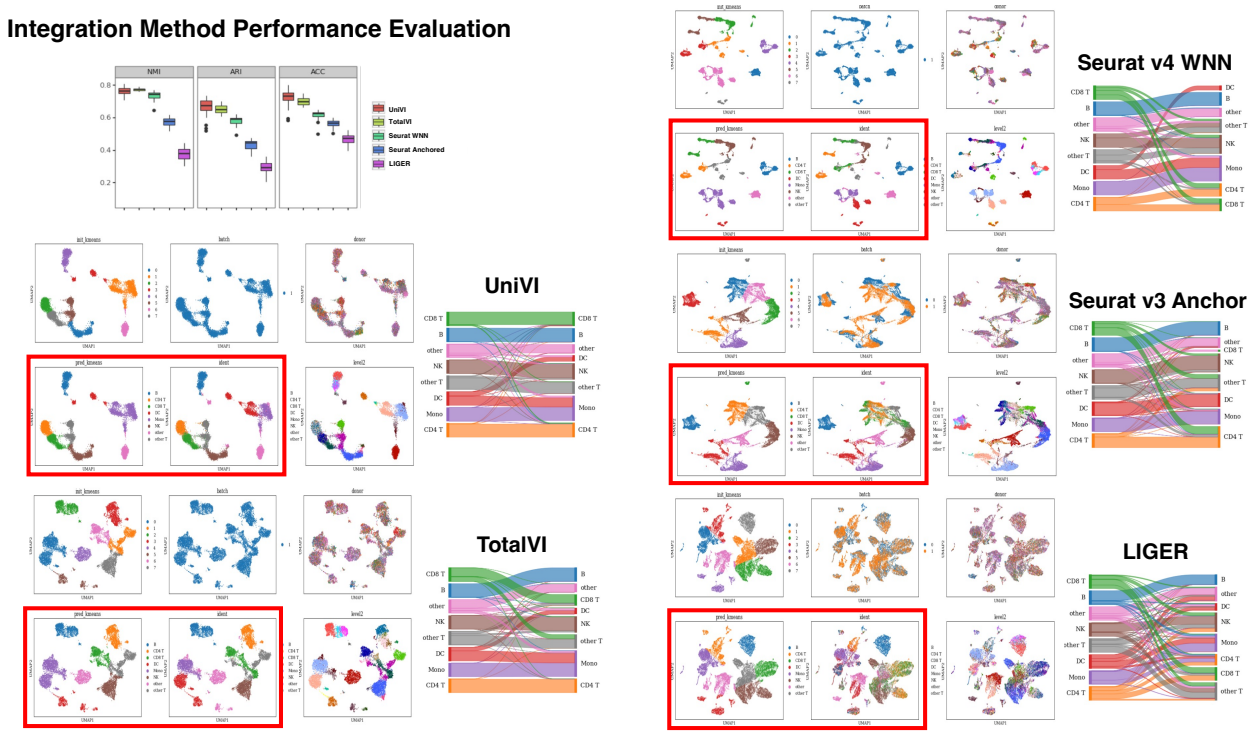

Fig. 4. UniVI Outperforms Existing State-of-the-Art Integration Methods. We underscore UniVI's exceptional performance when it comes to learning complex multimodal single-cell relationships between modalities from multimodal sequencing data. We compared our method to several commonly-used integration approaches and calculated metrics to measure different clustering qualities within the integrated data space. (*Top Left*) Integration performance metrics across 20 different replicates are illustrated using box plots for each integration method. The metrics evaluated include NMI, ARI, and ACC. NMI assesses the shared information between two clusterings, indicating how much information one clustering provides about the other. In this metric, UniVI outperforms all methods except TotalVI, likely due to TotalVI's utilization of prior information during data integration, whereas UniVI operates in a fully prior-agnostic manner. Additionally, TotalVI is a CITE-seq-specific method and cannot be generalized. For ARI, which measures the similarity between two clusterings while adjusting for chance groupings, UniVI surpasses all competing methods. Similarly, UniVI achieves the highest ACC, reflecting the proportion of correctly predicted cluster assignments out of the total data points. (*Rest of Figure*) Further evaluation of integration methods is conducted by comparing predicted supervised K-means clustering labels to actual cell type annotations using several different visualizations, including UMAPs colored by different variables and Sankey plots of the predicted versus actual cell types. UniVI outperforms the other four methods in aligning predicted clusters with true cell identities. Interestingly, a substantial proportion of cells annotated as Dendritic Cells (DC) are incorrectly clustered as Monocytes (Mono), a misclassification trend that is present across data integration methods to varying degrees.

and strong cross-modal alignment without requiring modality-specific priors, UniVI establishes itself as a versatile tool for scalable and interpretable multimodal data integration. Importantly, the ability to achieve high integration accuracy while maintaining biological fidelity makes UniVI well-suited for analyzing complex cellular systems, where multimodal integration can provide critical insights into cell identity, lineage relationships, and functional heterogeneity.

Expanding UniVI's Reach: Application to 10x Genomics Multiome Data

To further evaluate UniVI's versatility, we assessed its performance using single-cell multimodal data generated with the 10x Genomics Multiome platform. Unlike CITE-seq, which captures highly correlated transcriptomic and protein measurements, 10x Multiome assays profile scRNA-seq and scATAC-seq simultaneously. This presents a distinct computational challenge: while scRNA-seq captures gene expression, scATAC-seq measures chromatin accessibility, introducing fundamental differences in feature representation and data sparsity between modalities. Unlike the protein-based features in CITE-seq, which are relatively low-dimensional and directly linked to gene expression, scATAC-seq data is inherently high-dimensional, sparse, and less directly correlated with transcriptomic readouts. Thus, integrating these two modalities into a biologically coherent latent space requires methods that can effectively structure multimodal relationships despite these differences.

We trained UniVI on Dataset 2 (Materials and Methods) and assessed its capacity for multimodal integration, cross-modal imputation, and biologically informed denoising. Additionally, we tested UniVI's ability to generalize beyond directly paired datasets by applying a model trained on 10x Multiome (Dataset 3) to harmonize independent unimodal human PBMC datasets (Dataset 4: scRNA-seq; Dataset 5: scATAC-seq) (Materials and Methods). This experiment assessed whether UniVI could structure unimodal datasets within a shared latent space, demonstrating its utility for large-scale, atlas-level integration where direct modality pairings are unavailable.

To systematically assess UniVI's performance on 10x Multiome data, we conducted a series of experiments designed to evaluate its (1) ability to generalize across distinct multimodal sequencing technologies, (2) capacity to generate synthetic cross-modal samples and transfer learned cell identities across modalities, (3) effectiveness in harmonizing

disparate unimodal datasets within a multimodal framework, and (4) potential to mitigate technical noise and dropout effects in scRNA-seq data through multimodal-informed denoising. These analyses not only test UniVI's ability to integrate scRNA-seq and scATAC-seq data but also explore its broader applicability to single-cell data harmonization at multiple levels of granularity.

Validating UniVI's Robustness Across Diverse Data Sets

Integrating Dataset 2 using our method demonstrated strong alignment between scRNA-seq and scATAC-seq modalities, effectively mapping individual cells into a well-structured latent space with distinct clustering across major cell types. The integrated representations closely mirrored the caliber of the CITE-seq integration results, reinforcing UniVI's ability to generalize across multimodal sequencing technologies. However, challenges arose in downstream analyses due to imbalances in cell type representation, particularly affecting experiments reliant on deep feature learning, such as cross-modal classification.

Certain cell types, including $\gamma\delta$ T cells (gdT), mucosal-associated invariant T (MAIT) cells, naive B cells, plasmacytoid dendritic cells (pDC), hematopoietic stem and progenitor cells (HSPC), and plasma cells, were underrepresented, whereas others, such as CD4 Mono, CD4 Naive, CD8 Naive, and CD4 TCM, were overrepresented. The necessity of partitioning Dataset 2 into training and test sets further exacerbated this imbalance, limiting the model's ability to learn robust representations for rare cell populations. To mitigate this, we applied a sampling-with-replacement strategy during training, introducing additional variability to augment the learning of underrepresented populations. Preliminary results indicated that this approach improved representation depth, particularly in the cross-modally generated classification task.

Despite the potential risk of artificially clustered training data due to repeated sampling, we confirmed that this did not lead to overfitting, as unseen test samples remained well-integrated in the latent space (**Figure 5a**, Supplemental Figure 8). Before integration, test set cells exhibited clear modality-driven separation, with well-organized clusters corresponding to cell type identities. After integration, modalities aligned seamlessly, with clustering predominantly driven by biological variability rather than technical batch effects. Notably, while most cell types retained clear separation, some closely related clusters, such as CD4 Naive and CD8 Naive T cells, exhibited partial overlap, likely reflecting shared transcriptional programs and lineage transitions. The structured nature of the integrated space suggests that UniVI successfully captured both intra-cell type variation and biologically meaningful interrelationships between cell populations.

These findings highlight UniVI's capacity for robust multimodal data integration, preserving biologically relevant structure across diverse single-cell datasets. The well-clustered latent space is particularly crucial for UniVI's cross-generation capability, where data from one modality is encoded, sampled, and decoded into the alternate modality. Effective cross-generation requires both a well-defined latent space and a well-trained decoder, both of which were achieved through our integration and oversampling strategies. As larger single-cell datasets become available, these challenges are expected to be further mitigated, enabling even more refined UniVI models with deeper representations of cellular heterogeneity.

UniVI Effectively Transfers Cell Classes Across Modalities

To evaluate UniVI's ability to transfer learned information across modalities, we assessed its performance in cross-modal imputation using unseen data. A UniVI model trained on Dataset 2 was used to infer scATAC-seq chromatin accessibility profiles from scRNA-seq features in an independent test set (Materials and Methods). UMAP visualizations revealed that imputed scATAC-seq representations largely retained cell type identity, with reconstructed clusters aligning closely to expected differentiation trajectories. While some discrepancies between imputed and actual data were observed, the cross-modally generated data remained biologically coherent, with closely related cell types clustering in proximity, reflecting expected lineage transitions such as CD8 Naive to CD4 Naive to CD4 TCM. These results indicate that UniVI effectively preserves key biological features during cross-modal imputation, capturing both cell-state variation and lineage continuity.

To quantitatively assess the fidelity of imputed data, we trained classification neural networks on the cross-modally generated scATAC-seq features and evaluated their performance against real scATAC-seq data using the data unseen by our model (Materials and Methods). Two classification models were trained: one using a coarse annotation scheme (seven major cell types) and another using a high-resolution scheme (19 distinct cell subsets). The lower-resolution classifier performed robustly, with occasional misclassifications between CD4 and CD8 T cells, likely due to their shared transcriptomic and chromatin accessibility profiles. At higher granularity, misclassifications followed expected biological trends, particularly among transitional cell states such as CD4 Naive, CD4 TCM, and memory B cell subsets. These results suggest that UniVI maintains meaningful cell identity boundaries even when distinguishing closely related populations.

To further validate classification performance, we generated ROC AUC plots. At the broad immune lineage level, classification performance was highly accurate, with AUC values exceeding 0.98, indicating strong generalizability across major cell types. Certain populations, including B cells, monocytes, and natural killer cells, exhibited near-perfect classification (AUC > 0.999), consistent with their well-separated molecular profiles. At finer resolution, classification remained robust for most subsets, although CD4 Naive, CD8 TEM, and T regulatory (Treg) cells exhibited slightly lower AUC values, reflecting their transcriptional and chromatin accessibility similarities.

To ensure classification robustness, we performed 10-fold cross-validation, finding that performance remained stable across folds, with accuracy varying within approximately 5%. The average classification accuracy was 70.91% for high-resolution annotations and 87.19% for low-resolution annotations, with most classification errors occurring among biologically similar cell types rather than reflecting spurious predictions.

Overall, these findings confirm that UniVI preserves cell identity information across modalities, supporting high-fidelity imputation of missing or whole-sample data. Importantly, classification errors were generally constrained to biologically related cell types rather than arising from random noise, reinforcing UniVI's ability to retain meaningful cellular relationships within its learned latent space. These results

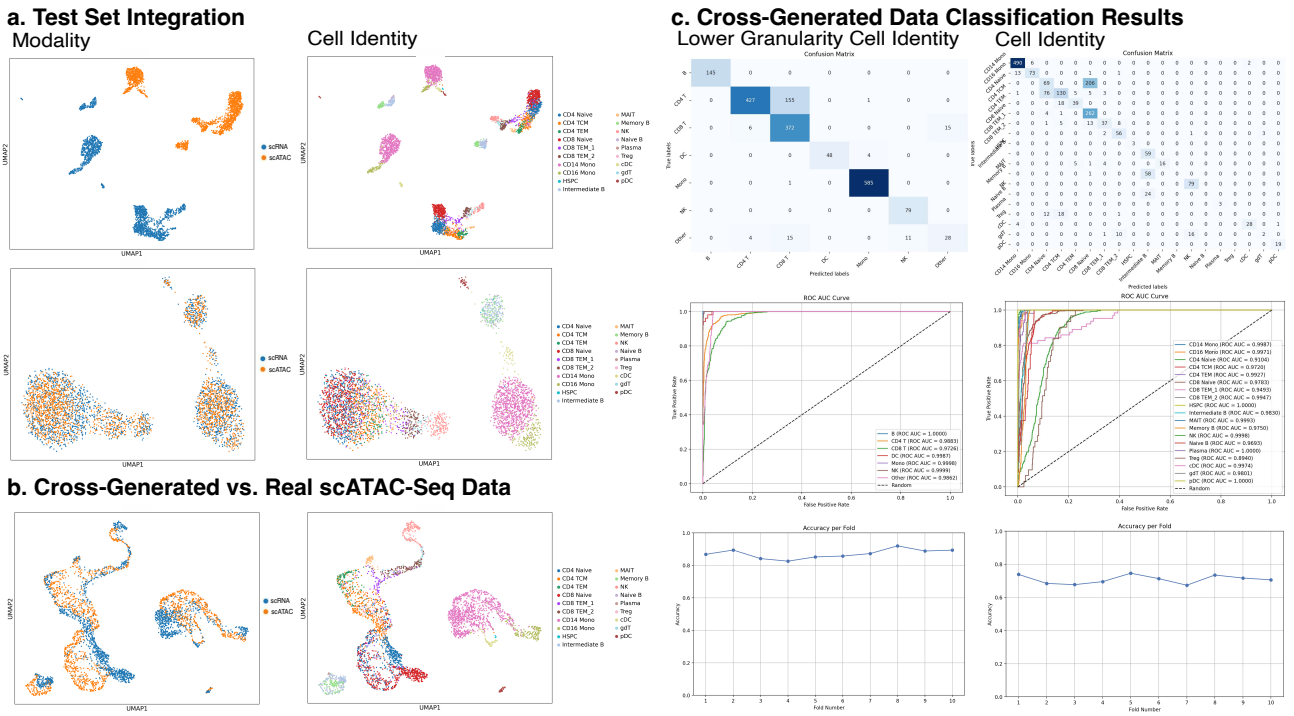

Fig. 5. Assessing UniVI Performance in Multimodal Single-Cell Analysis with Training and Cross-Modal Prediction Using Human PBMC Data. We show that UniVI is not only generalizable when it comes to integrating different multimodal sequencing technologies, but it has the ability to self—and cross—modally generate realistic samples that mimic real data. Here we visualize these results using 10x Multiome data. **(a.)** We first show that UniVI accurately integrates an unseen set of the data by showing the test set prior to—then after—integration. We plotted this separately colored by modality (Left), then colored by annotated cell type (Right). **(b.)** We highlight UMAPs of the scATAC-seq data that we reconstructed from scRNA-seq data (labeled as scRNA-seq in the modality-colored plot) and the actual scATAC-seq data. Moreover, we show the cross-generated data and the original data UMAP plot colored by cell identities, showing that cell identities are clustered together accurately inter—and intra—cluster-wise compared to the original data cell types. **(c.)** We next use the synthetic scATAC-seq samples to train two different classification neural networks—a low-resolution classifier with seven cell identity classes and a high-resolution classifier with 19 cell type classes. (*Top*) We highlight confusion matrices of both separate classification models. On the left, we show that the majority of the real scATAC-seq data are correctly classified outside of CD4 T cells, which struggled with classification due to unclear boundaries in the latent space between CD4 T cells and CD8 T cells. On the right, we showed that our high-granularity classification model performed well, except that it struggled to classify CD4 Naive T cells, misclassifying several of the cells as CD8 Naive T cells. We additionally see some confusion between different subgroups of B cells. This is again likely due to boundaries between groups that bleed together. (*Middle*) The ROC AUC curves show generally high-performance for the cross-modal prediction task, except again we see that the model struggles with the classification of a few classes as we previously discussed. (*Bottom*) Lastly, we show the classification accuracy for both models across 10-fold cross-validation. The plot representing the lower-resolution classifier performance showed that this model achieved 81% to 90% accuracy with about a 5% difference between folds. Whereas the high-resolution classifier showed between 68% to 75% accuracy. Again, the dips in accuracy are largely due to a few specific classes that likely do not have a clear learned boundary between them and the classes they are mistaken for. It would be interesting to see the performance using a model trained on more of these specific samples to see how well it is then able to distinguish them from one-another. Overall, we've shown that UniVI is broadly applicable to several multimodal sequencing technologies, it is also capable of generating biologically accurate cross-modal samples—depending on the granularity of the cell types and number of available samples for training.

underscore UniVI's potential for robust cross-modal data generation, demonstrating its utility for single-cell multimodal analysis by enabling accurate imputation and classification of missing modalities.

Pioneering Atlas-Level Integration Across Disparate Unimodal Datasets

To assess UniVI's capacity to integrate independent unimodal datasets, we applied a model trained on multimodal data to unify single-modality datasets with overlapping cell types. This experiment evaluates the feasibility of multimodal integration frameworks in harmonizing large-scale, atlas-level datasets, facilitating a more comprehensive understanding of cellular heterogeneity across independently generated single-cell datasets. For UniVI model training in this experiment, we used Dataset 3, a 10x Multiome human PBMC dataset with paired scRNA-seq and scATAC-seq, similar to the dataset used in the previous experiments but from a separate healthy donor (Materials and Methods). We then tested UniVI's ability to integrate two independent publicly available unimodal datasets: Dataset 4, an scRNA-seq dataset generated from multiple sequencing platforms, and Dataset 5, an scATAC-seq dataset (Materials and Methods).

We first evaluated the intra-modality harmonization of each dataset prior to full cross-modal alignment. Within Dataset 4, we observed significant batch effects arising from differences in sequencing technology, as the dataset was generated using eight distinct platforms. These inter-technology discrepancies, while expected, presented a challenge for dataset-wide integration. Despite the primary focus of this study being multimodal alignment, addressing batch effects within individual modalities remains a critical step

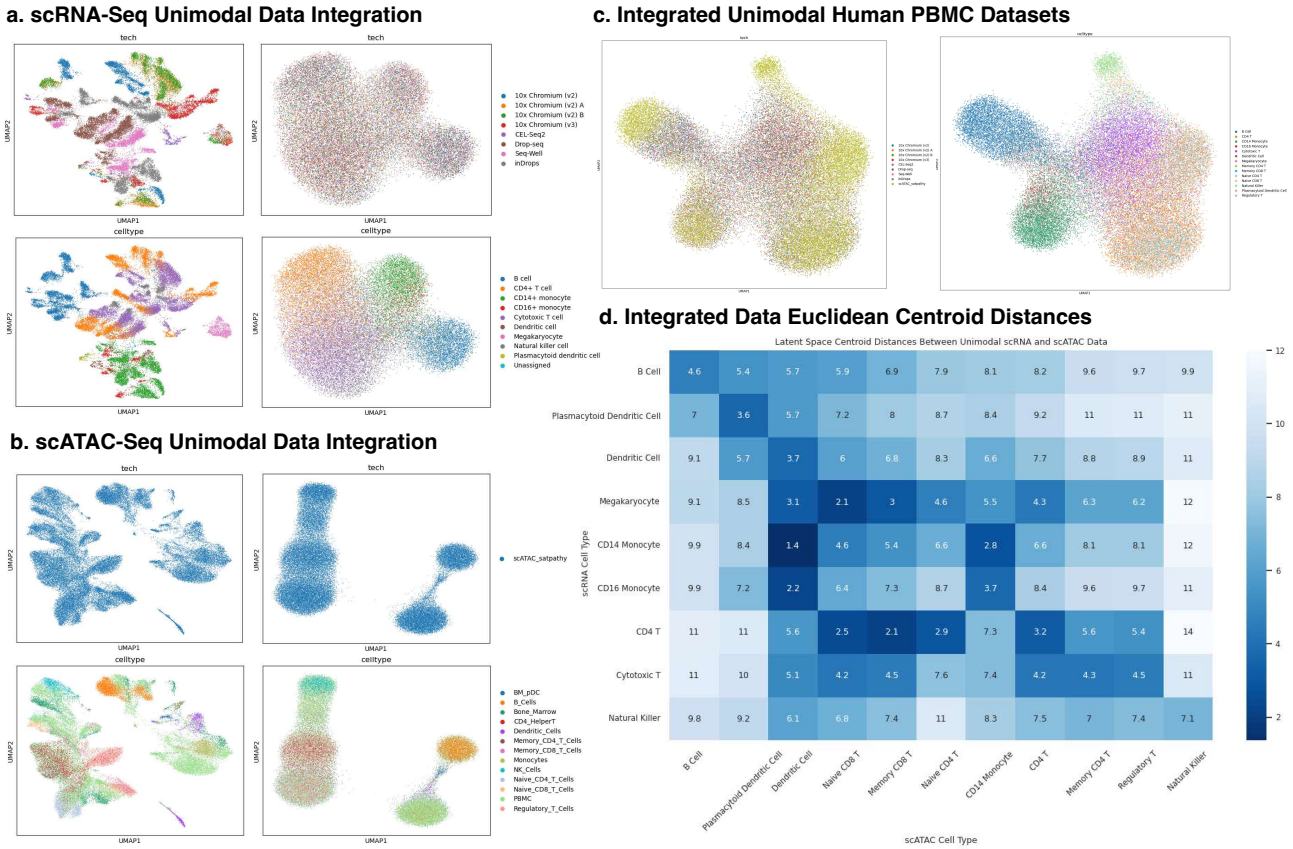

Fig. 6. Using UniVI to Integrate Separate Unimodal Datasets of Different Modalities to Enable Atlas-Level Data Integration. We spotlight UniVI's ability to integrate completely distinct unimodal datasets of different modalities of the same biological sample type, underscoring the potential to bring together a wealth of existing data. **(a.)** We first show just the unimodal scRNA-seq dataset prior to, and after, integration. (*Top*) We underscore the data by technology since the scRNA-seq data we used consists of eight different scRNA-seq technologies, which show significant batch effects between one-another. We spotlight that the within-modality integration largely eliminates the batch effects we saw prior to model application. (*Bottom*) We next show that within the scRNA-seq dataset, the integrated data space is mostly clustered by biological—as opposed to technical—variation. **(b.)** Next, we showed the scATAC-seq data prior to integration, then after. (*Top*) For consistency, we first plotted the scATAC-seq data by technology. We highlighted previously that several of the scRNA-seq samples were from different sequencing technologies, for consistency, we highlight that there is only one scATAC-seq technology. We especially underscore this since we later show the integration space colored by all technologies between both modalities. (*Bottom*) We next showed the same UMAP plot of the scATAC-seq data prior to integration, then after. We show here that the complex cell cluster-splitting artifacts are largely removed in the integrated space. We do show that there is over-dispersion in the scATAC-seq data after integration. Ultimately, we've reasoned that this phenomenon is an artifact of sequencing depth which is common within this modality, especially without reducing the dimensionality and discarding PC1 (see Results section for more details). **(c.)** Next, we highlight the integrated latent space between both unimodal sequencing modalities. (*Left*) When we show the integrated data space by technology, we show that there are several regions where all the technologies overlap one-another. However, there are some regions where the scATAC-seq data is clustered by itself, additionally it is pushed to the outer limits of the UMAP space. As previously mentioned, we've reasoned that this is largely an artifact of sequencing depth in the technology. (*Right*) Next, we show this same integrated space instead colored by cell type. This plot shows that similar cell types are clustered either overlapping one-another, or near each other between modalities. Of note, there are some cell types that are modality-specific. **(d.)** Finally, we calculated the Euclidean distance for all 60 latent variables between each modality, shown in a heatmap where the darker blue colors indicate a closer Euclidean distance between modalities. We separate these by cell types and see that between modalities, same—or similar—cell clusters are generally closest to one-another, indicating that the integrated space maintains biologically relevant information between modalities. Ultimately, this figure demonstrates UniVI's ability to integrate completely separate unimodal scRNA-seq and scATAC-seq datasets, effectively consolidating an "atlas-level" resource. Even when different batch effects and over-dispersion exist in each modality, UniVI's integrated latent space retains biologically meaningful cell type relationships across technologies and preserves the closest alignments among analogous cell subsets.

in ensuring robust downstream analyses. Post-integration, UniVI effectively mitigated these batch effects, with UMAP projections demonstrating clustering driven predominantly by biological identity rather than sequencing platform (**Figure 6a**). The model preserved meaningful cellular distinctions, with closely related populations remaining spatially proximate in the latent space, underscoring UniVI's ability to harmonize intra-modality variation while retaining biologically relevant structure.

We conducted a similar analysis for Dataset 5 (**Figure 6b**). Unlike Dataset 4, this dataset was derived from a single sequencing platform, reducing technical variability across samples. However, prior to integration, cell type clustering was less distinct with a notable fraction of samples assigned broad labels such as "Bone Marrow" and "PBMC," which likely obscured finer cellular distinctions. Following UniVI integration, cell type clusters exhibited improved separation, reflecting enhanced biological resolution within the dataset.

One challenge unique to the scATAC-seq dataset was the presence of substantial over-dispersion along UMAP1, a well-documented issue in chromatin accessibility data that arises due to variations in sequencing depth across individual cells [94]. This over-dispersion results in greater variance in feature detection, particularly in sparsely sequenced cells, contributing to artificial expansion of certain clusters. Prior studies have noted that scATAC-seq variability can be influenced by technical factors such as nucleosome positioning and sequencing depth [95], which may affect integration performance. Notably, when visualizing UMAP2 versus UMAP3, over-dispersion was substantially reduced (Supplemental Figure 9), suggesting that dimensionality reduction choices influence the observed distribution of scATAC-seq-derived embeddings. While some studies that reduce the dimensionality of their scATAC-seq data using singular value decomposition (SVD) propose the removal of the first principal component (PC1) to mitigate this effect [96], this approach risks discarding biologically meaningful variation. Additionally, dimensionality reduction in scATAC-seq data leaves researchers with an uninterpretable feature space. Given UniVI's ability to operate on both full—and reduced—dimensional representations (Supplemental Figure 10), future refinements in scATAC-seq preprocessing may further enhance integration outcomes.

Following unimodal dataset harmonization, we applied the 10x Multiome-trained UniVI model to map both unimodal datasets into a shared latent space, enabling direct cross-modality comparisons. To ensure consistency in cell type alignment, we harmonized annotations between Dataset 4 and Dataset 5, excluding ambiguous or poorly defined clusters (Materials and Methods). Post-integration, similar cell types across scRNA-seq and scATAC-seq modalities exhibited substantial overlap, highlighting UniVI's ability to align independent datasets based on shared biological characteristics (**Figure 6c**). However, some technical discrepancies persisted, particularly within the scATAC-seq modality, where certain clusters were positioned toward the periphery of the latent space. As observed in the unimodal analysis, this effect likely stems from sequencing depth variability rather than limitations in UniVI's integration framework. Alternative UMAP projections (Supplemental Figure 9) and dimensionality-reduced inputs (Supplemental Figure 10) further support this hypothesis, indicating that additional preprocessing steps tailored to scATAC-seq data could mitigate these residual batch effects.

To quantitatively assess integration fidelity, we computed Euclidean distances between cell type cluster centroids across all latent dimensions and visualized these relationships as a heatmap (**Figure 6d**). Expected cell type correspondences exhibited close proximity, confirming that UniVI accurately preserves biological relationships across modalities. However, as anticipated, rarer cell populations and those with ambiguous transcriptional profiles exhibited greater variability in alignment, suggesting potential avenues for further optimization in future multimodal integration workflows.

Overall, these findings establish UniVI's capacity to integrate single-cell datasets that lack direct multimodal pairing, demonstrating its broader applicability as a framework for large-scale atlas-level data harmonization. While challenges unique to scATAC-seq data remain, such as over-dispersion due to sequencing depth variability, UniVI effectively mitigates major batch effects and preserves biologically relevant cell type relationships across datasets. These results suggest that UniVI can serve as a powerful tool for bridging disparate single-cell datasets, facilitating cross-modal correlation studies and enabling integrative analyses that span multiple independent experimental platforms. As single-cell data resources continue to expand, incorporating additional dataset-specific preprocessing strategies and increasing training set diversity may further enhance UniVI's ability to generalize across diverse single-cell sequencing modalities.

UniVI Provides Robust, Biologically Relevant Noise-Reduction

Droplet-based scRNA-seq data is characterized by sparsity, largely due to limited sequencing depth per cell and stochastic transcript capture inefficiencies [66, 82]. This sparsity manifests as a high prevalence of zero counts, often attributed to the "dropout effect," where transcripts moderately expressed in one cell are absent in another, leading to data loss that must be accounted for in computational models that utilize scRNA-seq data [65, 66]. Given its prior success in denoising CITE-seq data, we sought to evaluate UniVI's ability to mitigate dropout effects and recover missing transcriptomic signals.

To systematically assess UniVI's denoising capabilities, we implemented a controlled dropout simulation inspired by the scVI manuscript [16]. Using a model trained on the Dataset 2 training set, we applied artificial dropout to the Dataset 2 unseen data while retaining the original data as ground truth (Materials and Methods). Dropout simulations were conducted at multiple levels (20%, 50%, and 70%) to reflect varying degrees of data sparsity that may be observed in real-world scRNA-seq experiments (and beyond). We compared heatmaps of the original scRNA-seq data with UniVI's reconstructions following dropout (**Figure 7a**), observing that UniVI effectively restores missing expression values and smooths noise, even at high dropout levels.

To quantify the accuracy of UniVI's reconstructions, we computed mean and median reconstruction errors for the artificially removed values (**Figure 7b**). UniVI demonstrated robust resilience to dropout-induced sparsity, with error rates remaining consistently low across all levels of simulated dropout. This suggests that the model effectively learns underlying gene expression patterns and can infer missing values based on shared latent representations.

Next, we examined whether UniVI's denoising preserves meaningful biological information by performing Gene Set Enrichment Analysis (GSEA) [97] on reconstructed data at dropout levels of 0%, 20%, 50%, and 70%. We used Kyoto Encyclopedia of Genes and Genomes (KEGG) [79, 98, 99] and Gene Ontology: Biological Process (GO: BP) [80, 100] pathway analyses to compare normalized enrichment scores (NESs) of reconstructed versus original datasets (**Figure 7c**). Pearson correlation coefficients revealed that UniVI-reconstructed datasets retained significantly stronger correlations with the original data compared to raw dropout simulations, confirming that UniVI effectively preserves biological signals even under extreme dropout conditions.

These results underscore UniVI's capacity to recover lost transcriptomic information and enhance data interpretability in scRNA-seq experiments. Given the persistent challenge of dropout in single-cell analysis, UniVI presents a powerful solution for mitigating data sparsity while preserving key biological relationships. Furthermore, while dropout is a known limitation primarily in scRNA-seq data, UniVI's denoising framework is broadly applicable across other modalities,

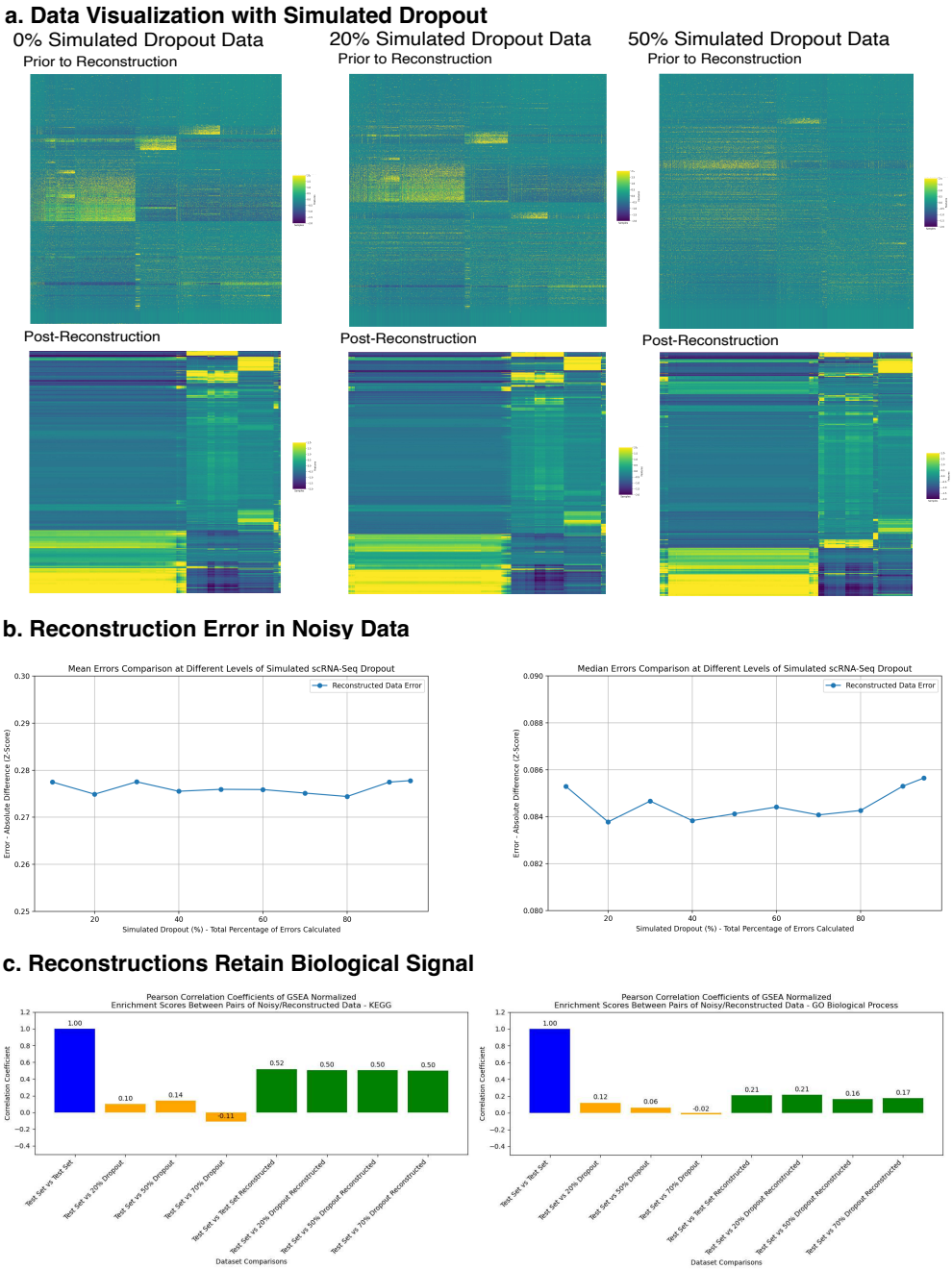

Fig. 7. UniVI Data Reconstructions Demonstrate Robustness to Noise and Preserve Biological Integrity. To assess UniVI's effectiveness in mitigating real-world technological artifacts, we simulated scRNA-seq data dropout at varying levels of data sparsity for counts above zero. **(a.)** Heatmaps illustrating original and reconstructed data under increasing dropout conditions. (*Left*) 0% dropout serves as the baseline with no data loss. (*Middle*) 20% dropout introduces moderate sparsity, showcasing UniVI's ability to effectively reconstruct missing values. (*Right*) 50% dropout represents severe data sparsity, highlighting UniVI's robust denoising performance under extreme conditions. **(b.)** Quantitative evaluation of UniVI's denoising capability through mean and median error rates between reconstructed and original data exclusively for the dropped values. (*Left*) Mean error rates remain stable across varying dropout levels, indicating consistent reconstruction accuracy. (*Right*) Median error rates demonstrate resilience to outliers, maintaining steady performance even at high dropout percentages. **(c.)** Assessment of biological signal preservation using Pearson correlation coefficients between NESs for KEGG and GO: Biological Processes gene sets. (*Left*) KEGG pathway correlations exhibit strong concordance between test set and reconstructed data across all dropout levels. (*Right*) GO Biological Processes correlations similarly demonstrate the retention of biological signals in reconstructed data despite increasing dropout percentages. Collectively, these results underscore UniVI's capability to robustly denoise noisy scRNA-seq data while preserving essential biological information.

offering a generalizable approach for enhancing the clarity and usability of multimodal single-cell datasets.

Discussion

UniVI demonstrates strong capabilities for integrating multimodal single-cell data without requiring explicit feature correspondences between modalities. Many traditional integration methods rely on predefined "anchors," such as gene-protein or gene-peak mappings, to align modalities, but these approaches can be restrictive when applied to datasets where such direct correspondences are unavailable or unreliable. By contrast, UniVI infers latent embeddings that encapsulate global relationships among cells, allowing for a more flexible and biologically coherent integration. This framework is particularly generalizable, as we've shown by unifying datasets generated from different sequencing technologies such as 10x Multiome and CITE-seq platforms, enabling seamless alignment across distinct measurement platforms. When benchmarked against established methods, UniVI consistently demonstrates high alignment accuracy, suggesting that its latent-factor approach effectively captures complex multimodal variation while preserving biologically meaningful relationships.

A key strength of UniVI lies in its use of a MoE VAE framework, which employs separate encoders for each modality and aligns them via a cross-latent space penalty. This structure allows the model to disentangle modality-specific variation while emphasizing biologically relevant signals shared across different data types. Unlike methods that rely on predefined correspondences or linear correlations, UniVI's nonlinear latent space can accommodate a wider range of biological phenomena, including sparse scATAC-seq peaks, asynchronous protein expression, and the dropout-prone nature of scRNA-seq data. By learning a shared latent representation that captures these complexities, UniVI achieves sharper clustering, more interpretable embeddings, and enhanced resolution of subtle cellular subpopulations.

Beyond integration, UniVI provides substantial advantages in denoising and imputation—critical capabilities for single-cell datasets prone to missing values. The model leverages its learned manifold constraints to infer missing features across modalities, generating biologically plausible reconstructions that enhance data interpretability. This extends beyond simple noise reduction; UniVI can generate entirely synthetic samples in one or more modalities, preserving the natural variability seen in real data. The ability to smooth transcriptomic and proteomic measurements improves cluster resolution and marker gene detection, ultimately benefiting downstream analyses such as differential expression and trajectory inference. Our results demonstrate that UniVI effectively reconstructs expression and accessibility patterns, further supporting its utility as a tool for multimodal single-cell analysis.

Despite its advantages, UniVI's performance is influenced by the quality of preprocessing and data distributional biases. The model assumes that the underlying distributions of each modality are sufficiently comparable. If left unaddressed, these biases can distort latent alignments, leading to systematic shifts that obscure true biological relationships. One strategy to mitigate these effects is the introduction of modality-aware correction mechanisms that dynamically adjust the contributions of each modality within the latent space. For instance, incorporating adaptive regularization terms could ensure that dominant modality-specific noise does not overshadow biologically relevant signals, ultimately enhancing integration robustness.

A particular challenge arises when integrating scATAC-seq data, due to the global chromatin accessibility profiles inherent to this modality. Unlike scRNA-seq, where gene expression is relatively localized and cell-type specific, chromatin accessibility is influenced by genome-wide regulatory landscapes and sequencing depth variability. These global accessibility patterns often become the dominant source of variation in dimensionality reduction techniques such as SVD, with sequencing depth artifacts overshadowing finer-scale biological variation. This over-dispersion effect, often captured in PC1 or the first dimension of a given dimensionality-reduction technique, can obscure meaningful biological signal and complicate multimodal alignment. Without properly accounting for these effects, latent space representations may be driven more by technical variation than true cellular differences, affecting the accuracy of downstream analyses.

Future iterations of UniVI could incorporate more sophisticated modeling approaches to account for these modality-specific biases. One promising direction is the use of adaptive weighting mechanisms that dynamically adjust the influence of scATAC-seq features based on inferred technical variation, preventing global accessibility effects from dominating the latent space. Additionally, the model could incorporate regulatory feature embeddings that explicitly disentangle sequencing depth artifacts from meaningful chromatin accessibility patterns, allowing for a more biologically relevant integration. By implementing such refinements, UniVI would maintain its advantages as a unified latent representation while improving its ability to capture cell type-specific chromatin landscapes with higher fidelity.

UniVI's modular architecture also enables the potential for seamless multimodal expansion. As sequencing technologies evolve to capture increasing numbers of data modalities, UniVI can be adapted to integrate additional features by extending its mixture-of-experts framework with additional encoders and decoders. This flexibility makes it well-suited for next-generation multimodal assays that measure transcriptomics, chromatin accessibility, spatial information, and epigenetic modifications in tandem. The ability to unify diverse data layers into a cohesive latent representation could pave the way for large-scale atlases of cellular states, enabling deeper insights into complex regulatory landscapes.

Beyond simultaneous multimodal integration, UniVI could be extended to align unimodal datasets from independent experiments. Currently, most integrative models require explicitly paired samples during training, limiting their applicability to datasets where such correspondences exist. However, a bipartite matching approach could allow UniVI to infer cross-modal relationships between independently trained unimodal datasets, effectively bridging datasets from different experimental conditions, tissues, or time points. A particularly intriguing use-case for this approach would be its potential applications to perturbation experiments. This would broaden UniVI's applicability, making it a more generalizable tool for harmonizing single-cell data across diverse biological contexts.

Ultimately, UniVI exemplifies how deep probabilistic modeling can be leveraged to unify disparate single-cell measurements into a coherent framework. Its ability to match or outperform state-of-the-art integration methods, coupled with its versatility in denoising, imputation, and synthetic data generation, establishes it as a powerful tool for single-cell analysis. As multimodal sequencing technologies continue

to advance, refinements to UniVI—such as improved bias correction, expanded multimodal architectures, and cross-dataset alignment—will further enhance its utility in bridging the gap between specialized assays and a more holistic understanding of cellular function. By integrating diverse single-cell data sources into a unified framework, UniVI has the potential to accelerate discoveries in cell biology, disease research, and precision medicine.

Availability of Code and Implementation

The UniVI framework, along with its source code, computational workflows, and user-friendly tutorials, is openly accessible at `https://github.com/Ashford-A/UniVI`. This repository includes comprehensive examples and documentation to facilitate reproducibility and adaptation of the model to other multimodal single-cell datasets. Users are encouraged to explore its applications across diverse biological contexts and contribute to its ongoing development.

Supplementary Information

Detailed supplementary data, including extended analyses, figures, and tables that support and expand upon the findings presented in this manuscript, are available online at *Nucleic Acids Research*. These supplementary materials are designed to provide a deeper insight into the methodologies and results discussed, offering a valuable resource for researchers aiming to replicate or extend the study.

Acknowledgments

The authors would like to extend sincere appreciation to Hyeyoung Cho for her foundational contributions to this project which were instrumental in its inception and development. The authors additionally express their gratitude to Oregon Health & Science University for computational support provided through the Exacloud cluster. They also acknowledge the contributors of Hao et al. and data platforms such as atlas.fredhutch.org and the Gene Expression Omnibus for making CITE-seq data publicly available for research purposes. Special thanks are extended to 10x Genomics for generating and hosting the 10x Multiome datasets integral to this study.

Declarations

The authors declare no conflicts of interest associated with this manuscript.

Funding

This work was supported by the International Alliance for Cancer Early Detection, an alliance between Cancer Research UK, Canary Center at Stanford University, the University of Cambridge, OHSU Knight Cancer Institute, University College London and the University of Manchester.

Equations

Equations 1, 2, 3, and 5 describe the core UniVI model components (Equation 4 describes the specific decoder equations used whereas Equation 3 describes the generalized decoder equation). Equations 6, 7, 8, and 9 detail the derivation of the UniVI loss function, culminating in the final form in Equation 9. Finally, Equations 10, 11, 12, and 13 describe performance metrics used in this manuscript to evaluate the integration performance of trained UniVI models. Specifically, Equations 11, 12, and 13 describe the qualities of the integrated space used in the integration tool performance experiments. Equation 10 was mostly used to quantify integration performance when comparing performance among UniVI models with different parameters, preprocessing methods, and feature selection steps.

Equation 1: Encoder Equation

$$q_{\phi_k}(z_k \mid x_k) = \mathcal{N}\Big(z_k;\, \mu_{\phi_k}(x_k),\, \sigma^2_{\phi_k}(x_k)\Big) \tag{1}$$

Equation 2: Prior Distribution Calculation

$$p(z_P) = \mathcal{N}\big(z_k;\, \mu_P,\, \Sigma_P\big) \tag{2}$$

Equation 3: Generalized Decoder Equation

$$p_{\theta_k}(x_k \mid z_k) = \begin{cases} \mathcal{N}\big(x_k;\, \mu_{\theta_k}(z_k),\, \sigma^2_{\theta_k}(z_k)\big), & \text{if data is continuous (Gaussian decoder)}, \\ \mathrm{NB}\big(x_k;\, r_{\theta_k}(z_k),\, p_{\theta_k}(z_k)\big), & \text{if data is count-based (Negative Binomial decoder)} \end{cases} \tag{3}$$

Equation 4: Specific Decoder Implementations

a. Gaussian Decoder:

$$\mu = W z_k + b, \qquad \sigma^2 = \exp\big(\mathrm{Softplus}(W z_k + b)\big) + \eta, \quad \text{where } \mathrm{Softplus}(x) = \log\big(1 + e^x\big) \tag{4a}$$

b. Negative Binomial Decoder:

$$r = \min\Big(\mathrm{Softplus}(W z_k + b),\ \mathrm{clamp_max}\Big), \qquad p = \mathrm{clamp}\Big(\sigma(W z_k + b),\ \epsilon,\ 1 - \epsilon\Big), \quad \text{where } \sigma(x) = \frac{1}{1 + e^{-x}} \tag{4b}$$

Equation 5: ReLU Activation Function

$$\mathrm{ReLU}(x) = \begin{cases} x, & \text{if } x > 0, \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

Equation 6: Mixture of Experts Variational Autoencoder ELBO

$$L_{\text{MoE VAE ELBO}} = \sum_{k=1}^{K} \Big[\mathbb{E}_{q_{\phi_k}(z_k \mid x_k)} \big[\log p_{\theta_k}(x_k \mid z_k)\big] \;-\; D_{KL}\big(q_{\phi_k}(z_k \mid x_k) \,\big\|\, p(z_P)\big) \Big] \tag{6}$$

Equation 7: β-VAE Term Addition

$$L_{\text{MoE } \beta\text{-VAE ELBO}} = \sum_{k=1}^{K} \Big[\mathbb{E}_{q_{\phi_k}(z_k \mid x_k)} \big[\log p_{\theta_k}(x_k \mid z_k)\big] \;-\; \beta \cdot D_{KL}\big(q_{\phi_k}(z_k \mid x_k) \,\big\|\, p(z_P)\big) \Big] \tag{7}$$

Equation 8: Additional Cross-Modal KL Divergence

$$L_{\text{MoE } \beta\text{-VAE plus additional}} = \sum_{k=1}^{K} \Big[\mathbb{E}_{q_{\phi_k}(z_k \mid x_k)} \Big[\sum_{j=1}^{K} \log p_{\theta_{kj}}(x_j \mid z_k) \Big] \;-\; \beta \cdot D_{KL}\big(q_{\phi_k}(z_k \mid x_k) \,\big\|\, p(z_P)\big) \;-\; \beta \sum_{\substack{j=1 \\ j \neq k}}^{K} D_{KL}\big(q_{\phi_k}(z_k \mid x_k) \,\big\|\, q_{\phi_j}(z_j \mid x_j)\big) \Big] \tag{8}$$

Equation 9: Final UniVI Loss Function

$$L_{\text{UniVI}} = \sum_{k=1}^{K} \Big[\mathbb{E}_{q_{\phi_k}(z_k \mid x_k)} \Big[\sum_{j=1}^{K} \log p_{\theta_{kj}}(x_j \mid z_k) \Big] \;-\; \beta \cdot D_{KL}\big(q_{\phi_k}(z_k \mid x_k) \,\big\|\, p(z_P)\big) \;-\; \gamma \sum_{\substack{j=1 \\ j \neq k}}^{K} D_{KL}\big(q_{\phi_k}(z_k \mid x_k) \,\big\|\, q_{\phi_j}(z_j \mid x_j)\big) \Big] \tag{9}$$

Equation 10: Fraction of Samples Closer Than the True Match (FOSCTTM)

$$\text{FOSCTTM}_i = \frac{\sum_{k \neq j} \mathbb{I}\big(d(i,k) < d(i,j)\big)}{N_{\text{modality 2}}} \quad \text{where } d(i,k) = \sqrt{\sum_{m=1}^{M} \left(x_i^{(m)} - x_k^{(m)}\right)^2} \tag{10}$$

Equation 11: Normalized Mutual Information (NMI)

$$\text{NMI}(U,V) = \frac{2\, I(U;\, V)}{H(U) \;+\; H(V)} \tag{11}$$

Equation 12: Adjusted Rand Index (ARI)

$$\text{ARI} = \frac{\sum_{i,j} \binom{n_{ij}}{2} \;-\; \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} \;+\; \sum_j \binom{b_j}{2}\right] \;-\; \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}} \tag{12}$$

Equation 13: Clustering Accuracy (ACC)

$$\text{ACC} = \frac{1}{n} \sum_{i=1}^{n} \delta\big(\text{true label}_i,\ \text{map}(\text{cluster label}_i)\big) \tag{13}$$

Appendix A Publication Experiment Reproducibility

A.1 Algorithm Pseudocode

Algorithm 1 UniVI Training Algorithm

```
Initialize variational parameters φ_k for each expert k
Initialize generative parameters θ_k for each expert k
Initialize best validation loss best_val_loss ← ∞
Initialize patience counter patience ← 0
Set patience threshold patience_threshold
for each epoch do
    for each batch {x_k}_{k=1}^K from modalities 1 to K do
        for each modality k do
            Encode x_k into latent space: z_k ~ q_{φ_k}(z_k|x_k)
            Compute expected log-likelihood for modality k: E_{q_{φ_k}(z_k|x_k)} [Σ_{j=1}^K log p_{θ_{k_j}}(x_j|z_k)]
        end for
        Initialize total loss: L_total = 0
        for each modality k do
            Compute within-modality KL divergence: L_KL^k = D_KL(q_{φ_k}(z_k|x_k)||p(z_P))
            Add to total loss: L_total += β L_KL^k
            for each modality j ≠ k do
                Compute cross-modal KL divergence: L_KL^{kj} = D_KL(q_{φ_k}(z_k|x_k)||q_{φ_j}(z_j|x_j))
                Add to total loss with weight: L_total += γ L_KL^{kj}
            end for
        end for
        Update {φ_k} and {θ_k} for all experts to minimize L_total
    end for
    Evaluate model on validation set and get val_loss
    if val_loss < best_val_loss then
        best_val_loss ← val_loss
        patience ← 0
        Save model parameters as the best so far
    else
        patience ← patience + 1
    end if
    if patience ≥ patience_threshold then
        break                                          ▷ Early stopping triggered
    end if
end for
```

A.2 Model Hyperparameters for CITE-Seq Data Training

Note: The same hyperparameters were applied across all CITE-seq experiments.

- **Learning Rate:** Set at 0.001 to facilitate smooth and gradual learning.
- **Optimizer:** Adam optimizer, selected for its ability to handle sparse gradients effectively and adjust learning rates dynamically.
- **Batch Size:** Set to 500 to optimize computational efficiency while maintaining training stability.
- **Number of Epochs:** Training is regulated by monitoring validation loss, with early stopping to prevent overfitting once convergence is achieved.
- **Latent Space Dimension (z):** Set to 50 to adequately capture complex data structures while fitting the nature of CITE-seq data.
- **Number of Hidden Layers:** Configured with 3 layers for Modality 1 (RNA) and 2 layers for Modality 2 (ADT), reflecting the larger number of features in scRNA-seq data compared to scADT (2,000 versus 224 features).
- **β-Value in Loss Function:** Fixed at 100, ensuring the latent manifolds align during training and controlling disentanglement in the latent space.
- **γ-Value in Loss Function:** Set to 100, which was optimal to learn alignment the two modalities during training.
- **Number of Monte Carlo Samples:** 20 samples are used to approximate expectations during the training process.
- **Dimensions of Hidden Layers:** Set at 128 to balance the model's complexity and its computational feasibility, recognizing the reduced complexity of protein data compared to scRNA-seq data.
- **Number of scRNA-seq Input Features (X1):** The top 2,000 highly variable transcripts are selected to capture a broad spectrum of biological signals.
- **Number of scADT Input Features (X2):** All 224 scADT features are included due to the relatively small size of the feature set compared to the scRNA-seq feature set.

A.3 Model Hyperparameters for 10x Multiome Data Training

This configuration is used for visualizing the latent space and training a classifier on scRNA-generated ATAC data.

- **Learning Rate:** Set to 0.001, ensuring smooth gradient updates and stable learning.
- **Optimizer:** The Adam optimizer is chosen for its efficiency in dealing with sparse gradients and its ability to adapt learning rates.
- **Batch Size:** Set at 500, balancing computational efficiency with model stability.
- **Number of Epochs:** Determined by monitoring the convergence of validation loss, with early stopping to avoid overfitting.
- **Latent Space Dimension (z):** Fixed at 60 to capture the necessary complexity of data representation while effectively reducing feature dimensionality.
- **Number of Hidden Layers:** Configured with 3 layers for scRNA-seq (modality 1) and 5 layers for ATAC (modality 2), to accommodate the higher complexity of ATAC data.
- **β-Value in Loss Function:** Set at 145 to manage latent space disentanglement and prioritize the penalty on cross-modal KL-divergence.
- **γ-Value in Loss Function:** Set to 145, which allowed the model to learn adequate cross-modal latent representations.
- **Number of Monte Carlo Samples:** 20 samples are used to ensure sufficient approximation of expectations during training.
- **Dimensions of Hidden Layers:** Set at 224 to achieve a balance between the complexity of the model and its computational requirements.
- **Number of scRNA-seq Input Features (X1):** The top 2,000 highly variable transcripts are selected to capture a comprehensive array of biological signals while maintaining efficiency.
- **Number of ATAC Input Features (X2):** The top 2,000 highly variable peaks are selected, optimizing the model's ability to differentiate between cells.

Model Hyperparameters for Outside Uni-Modal Data Integration Experiment Training

These parameters were used for the Integrating Disparate Unimodal Datasets using Pre-Trained UniVI Model experiment.

- **Learning Rate:** Set to 0.001 to ensure gradual adjustments during training.
- **Optimizer:** Adam optimizer, chosen for its capability to handle sparse gradients effectively and adjust learning rates adaptively.
- **Batch Size:** Set to 500 to strike a balance between computational efficiency and training stability.
- **Number of Epochs:** The training process continues until a steady state is reached in validation loss, with early stopping employed to prevent overfitting.
- **Latent Space Dimension (z):** Fixed at 60, enabling the model to encapsulate complex data patterns while sufficiently compressing feature dimensions, especially relevant for 10x Multiome data.
- **Number of Hidden Layers:** Configured with 3 hidden layers for scRNA-seq (modality 1) and 5 for ATAC (modality 2), reflecting the higher complexity of ATAC data.
- **β-Value in Loss Function:** Set to 145, to encourage disentangled representation in the latent space, emphasizing cross-modal KL-divergence penalties.
- **γ-Value in Loss Function:** Set to 145, which allowed the model to learn adequate cross-modal latent representations for this specific data type.
- **Number of Monte Carlo Samples:** Using 20 samples ensures an adequate approximation of expectations during the training process.
- **Dimensions of Hidden Layers:** Set to 328 to maintain a balance between adequately modeling the data and keeping computational requirements manageable.
- **Number of scRNA-seq Input Features (X1):** Limited to the top 2,000 highly variable transcripts, ensuring a broad capture of biological signals while keeping the computational load reasonable.
- **Number of ATAC Input Features (X2):** Top 5,000 highly variable peaks are selected due to the higher volume of ATAC measurements per cell and the sparse nature of the data, providing sufficient feature diversity for cell differentiation.

Model Hyperparameters for Dropout Effect Experiments Training

These parameters were used for the Denoising Experiments using Simulated Data Dropout experiment.

- **Learning Rate:** Initiated at 0.001 to facilitate smooth gradient updates throughout the training.
- **Optimizer:** The Adam optimizer is employed for its efficiency in managing sparse gradients and dynamically adjusting learning rates.
- **Batch Size:** Set to 500, aiming to optimize the use of computational resources.
- **Number of Epochs:** Training duration is determined by monitoring validation loss, with early stopping applied to avoid overfitting.
- **Latent Space Dimension (z):** Chosen to be 60, providing a balanced approach to capturing the necessary complexity of the data while reducing feature dimensionality.
- **Number of Hidden Layers:** Configured with 3 layers for scRNA-seq data and 5 for ATAC data, reflecting the varying complexity levels across these modalities.
- **β-Value in Loss Function:** A value of 200 is used to control the degree of latent space disentanglement, with a focus on cross-modal latent feature penalties.
- **γ-Value in Loss Function:** Set to 200, which allowed the model to learn adequate cross-modal latent representations for these specific experiments.
- **Number of Monte Carlo Samples:** Using 20 samples for a sufficient approximation of expectations during the training phases.
- **Dimensions of Hidden Layers:** Set at 224, offering a balance between model complexity and computational efficiency.
- **Number of scRNA-seq Input Features (X1):** The top 2,000 highly variable transcripts are chosen, capturing a wide range of gene expressions while maintaining efficiency.
- **Number of ATAC Input Features (X2):** Limited to the top 2,000 highly variable peaks, streamlining the feature set to enhance model performance in identifying key differences between cells.

A.4 Early Stopping

Early stopping was set to 20 epochs, which seemed to be an adequate patience value for all experiments performed.

References

1. I. C. Macaulay, W. Haerty, P. Kumar, Y. I. Li, T. X. Hu, M. J. Teng, M. Goolam, N. Saurat, P. Coupland, L. M. Shirley, M. Smith, N. Van der Aa, R. Banerjee, P. D. Ellis, M. A. Quail, H. P. Swerdlow, M. Zernicka-Goetz, F. J. Livesey, C. P. Ponting, and T. Voet. G&t-seq: Parallel sequencing of single-cell genomes and transcriptomes. *Nature Methods*, 12:519–522, 2015.
2. C. Angermueller, S. J. Clark, H. J. Lee, I. C. Macaulay, M. J. Teng, T. X. Hu, F. Krueger, S. A. Smallwood, C. P. Ponting, T. Voet, G. Kelsey, O. Stegle, and W. Reik. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods*, 13:229–232, 2016.
3. A. Ma, A. McDermaid, J. Xu, Y. Chang, and Q. Ma. Integrative methods and practical challenges for single-cell multi-omics. *Trends in Biotechnology*, 38(9):1007–1022, 2020.
4. A. Baysoy, Z. Bai, R. Satija, and R. Fan. The technological landscape and applications of single-cell multi-omics. *Nature Reviews Molecular Cell Biology*, 24:695–713, 2023.
5. C. Marr, J. X. Zhou, and S. Huang. Single-cell gene expression profiling and cell state dynamics: collecting data, correlating data points and connecting the dots. *Current Opinion in Biotechnology*, 39:207–214, 2016.
6. L. Haghverdi, A. T. L. Lun, M. D. Morgan, and J. C. Marioni. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36:421–427, 2018.
7. T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. III Mauck, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.e21, 2019.
8. A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36:411–420, 2018.
9. I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P. Loh, and S. Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature Methods*, 16:1289–1296, 2019.
10. D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16:2639–2664, 2004.
11. G. Peyré and M. Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.
12. Z.-J. Cao and G. Gao. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 40:1458–1466, 2022.
13. Z. He, S. Hu, Y. Chen, S. An, J. Zhou, R. Liu, J. Shi, J. Wang, G. Dong, J. Shi, J. Zhao, L. Ou-Yang, Y. Zhu, X. Bo, and X. Ying. Mosaic integration and knowledge transfer of single-cell multimodal data with midas. *Nature Biotechnology*, 42:1594–1605, 2024.
14. K. Minoura, K. Abe, H. Nam, H. Nishikawa, and T. Shimamura. A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. *Cell Reports Methods*, 1(5):100071, 2021.
15. S. G. Stark, J. Ficek, F. Locatello, X. Bonilla, S. Chevrier, F. Singer, Tumor Profiler Consortium, G. Rätsch, and K.-V. Lehmann. Scim: Universal single-cell matching with unpaired feature sets. *Bioinformatics*, 36(Supplement_2):i919–i927, 2020.
16. R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15:1053–1058, 2018.
17. K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang, and X. Zhuang. Spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233):aaa6090, 2015.
18. H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.
19. J. Dou, S. Liang, V. Mohanty, Q. Miao, Y. Huang, Q. Liang, X. Cheng, S. Kim, J. Choi, Y. Li, L. Li, M. Daher, R. Basar, K. Rezvani, R. Chen, and K. Chen. Bi-order multimodal integration of single-cell data. *Genome Biology*, 23(112), 2022.

20. I. González, S. Déjean, P. G. P. Martin, O. Gonçalves, P. Besse, and A. Baccini. Highlighting relationships between heterogeneous biological data through graphical displays based on regularized canonical correlation analysis. *Journal of Biological Systems*, 17(2):173–199, 2009.
21. E. Parkhomenko, D. Tritchler, and J. Beyene. Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 8(1):Article 1, 2009.
22. C. Soneson, H. Lilljebjörn, T. Fioretos, and M. Fontes. Integrative analysis of gene expression and copy number alterations using canonical correlation analysis. *BMC Bioinformatics*, 11:Article 191, 2010.
23. S. Waaijenborg, P. C. Verselewel de Witt Hamer, and A. H. Zwinderman. Quantifying the association between gene expressions and dna-markers by penalized canonical correlation analysis. *Statistical Applications in Genetics and Molecular Biology*, 7:Article 3, 2008.
24. D. M. Witten and R. J. Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, 8:Article 1, 2009.
25. S. Chen, B. Zhu, S. Huang, J. W. Hickey, K. Z. Lin, M. Snyder, W. J. Greenleaf, G. P. Nolan, N. R. Zhang, and Z. Ma. Integration of spatial and single-cell data across modalities with weakly linked features. *Nature Biotechnology*, 42:1096–1106, 2024.
26. C. Bunne, S. G. Stark, G. Gut, J. Sarabia del Castillo, M. Levesque, K.-V. Lehmann, L. Pelkmans, A. Krause, and G. Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20:1759–1768, 2023.
27. K. Cao, Q. Gong, Y. Hong, and L. Wan. A unified computational framework for single-cell data integration with optimal transport. *Nature Communications*, 13:7419, 2022.
28. P. Demetci, R. Santorella, B. Sandstede, W. S. Noble, and R. Singh. Scot: Single-cell multi-omics alignment with optimal transport. *Journal of Computational Biology*, 29(1), 2022.
29. G.-J. Huizing, G. Peyré, and L. Cantini. Optimal transport improves cell–cell similarity inference in single-cell omics data. *Bioinformatics*, 38(8):2169–2177, 2022.
30. N. Moriel, E. Senel, N. Friedman, N. Rajewsky, N. Karaiskos, and M. Nitzan. Novosparc: Flexible spatial reconstruction of single-cell gene expression with optimal transport. *Nature Protocols*, 16:4177–4200, 2021.
31. S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren. Multimodal deep learning for biomedical data fusion: A review. *Briefings in Bioinformatics*, 23(2), 2022.
32. C. Xiao, Y. Chen, Q. Meng, L. Wei, and X. Zhang. Benchmarking multi-omics integration algorithms across single-cell rna and atac data. *Briefings in Bioinformatics*, 25(2), 2024.
33. D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
34. I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. Beta-VAE: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.
35. Y. Shi, N. Siddharth, B. Paige, and P. H. S. Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
36. Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 322–330, 2019.
37. C. Wang and S. Mahadevan. A general framework for manifold alignment. In *Proceedings of the 24th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 1298–1306, 2009.
38. L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2020.
39. H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441, 498–520, 1933.
40. I. T. Jolliffe. *Principal component analysis*. Springer, New York, NY, 2nd edition, 2002.
41. M. Stoeckius, C. Hafemeister, W. Stephenson, B. Houck-Loomis, P. K. Chattopadhyay, H. Swerdlow, R. Satija, and P. Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14:865–868, 2017.
42. J. D. Buenrostro, B. Wu, H. Y. Chang, and W. J. Greenleaf. Atac-seq: A method for assaying chromatin accessibility genome-wide. *Current Protocols in Molecular Biology*, 109:21.29.1–21.29.9, 2015.
43. 10x Genomics. 10k human pbmcs, multiome v1.0, chromium x, single-cell multiome atac + gene expression dataset generated by a chromium x series instrument, analyzed using cell ranger arc 2.0.0, 2021. Available online.
44. 10x Genomics. Pbmc from a healthy donor - no cell sorting (10k), single-cell multiome atac + gene expression dataset generated by a chromium controller instrument and analyzed by cell ranger arc 2.0.0, 2021. Available online.
45. Y. Hao, S. Hao, E. Andersen-Nissen, W. M. III Mauck, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, P. Hoffman, M. Stoeckius, E. Papalexi, E. P. Mimitou, J. Jain, A. Srivastava, T. Stuart, L. M. Fleming, B. Yeung, A. J. Rogers, and R. Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, 2021.
46. M. L. Elizaga, S. S. Li, N. K. Kochar, G. J. Wilson, M. A. Allen, H. V. N. Tieu, I. Frank, M. E. Sobieszczyk, K. W. Cohen, B. Sanchez, T. E. Latham, D. K. Clarke, M. A. Egan, and NIAID HIV Vaccine Trials Network (HVTN) 087 Study Team. Safety and tolerability of hiv-1 multiantigen pdna vaccine given with il-12 plasmid dna via electroporation, boosted with a recombinant vesicular stomatitis virus hiv gag vaccine in healthy volunteers in a randomized, controlled clinical trial. *PLOS ONE*, 13(9):e0202753, 2018.
47. S. S. Li, N. K. Kochar, M. Elizaga, C. M. Hay, G. J. Wilson, K. W. Cohen, S. C. De Rosa, and NIAID HIV Vaccine Trials Network. Dna priming increases frequency of t-cell responses to a vesicular stomatitis virus hiv vaccine with specific enhancement of cd8+ t-cell responses by interleukin-12 plasmid dna. *Clinical and Vaccine Immunology*, 24(11), 2017.

48. J. Ding, X. Adiconis, S. K. Simmons, M. S. Kowalczyk, C. C. Hession, N. D. Marjanovic, T. K. Hughes, M. H. Wadsworth, T. Burks, L. T. Nguyen, J. Y. H. Kwon, B. Barak, W. Ge, A. J. Kedaigle, S. Carroll, S. Li, N. Hacohen, O. Rozenblatt-Rosen, A. K. Shalek, A.-C. Villani, A. Regev, and J. Z. Levin. Systematic comparison of single-cell and single-nucleus rna-sequencing methods. *Nature Biotechnology*, 38:737–746, 2020.
49. A. T. Satpathy, J. M. Granja, K. E. Yost, Y. Qi, F. Meschi, G. P. McDermott, B. N. Olsen, M. R. Mumbach, S. E. Pierce, M. R. Corces, P. Shah, J. C. Bell, D. Jhutty, C. M. Nemec, J. Wang, L. Wang, Y. Yin, P. G. Giresi, A. L. S. Chang, G. X. Y. Zheng, W. J. Greenleaf, and H. Y. Chang. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nature Biotechnology*, 37:925–936, 2019.
50. Fred Hutchinson Cancer Research Center. Multimodal PBMC dataset from HIV vaccine trial donors, 2021.
51. Gene Expression Omnibus. Dataset GSE164378, 2021.
52. Database of Genotypes and Phenotypes. Dataset phs002315.v1.p1, 2021.
53. I. Virshup, S. Rybakov, F. J. Theis, P. Angerer, and F. A. Wolf. anndata: Annotated data. *bioRxiv*, 2021.
54. F. A. Wolf, P. Angerer, and F. J. Theis. Scanpy: Large-scale single-cell gene expression data analysis. *Genome Biology*, 19:15, 2018.
55. X. Lin, T. Tian, Z. Wei, and H. Hakonarson. Clustering of single-cell multi-omics data with a multimodal deep learning method. *Nature Communications*, 13:7705, 2022.
56. R. Argelaguet, D. Arnol, D. Bredikhin, Y. Deloro, B. Velten, J. C. Marioni, and O. Stegle. Mofa+: A statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, 21:111, 2020.
57. J. D. Welch, V. Kozareva, A. Ferreira, C. Vanderburg, C. Martin, and E. Z. Macosko. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, 177(7):1873–1887.e17, 2019.
58. L. Xiong, K. Tian, Y. Li, W. Ning, X. Gao, and Q. C. Zhang. Online single-cell data integration through projecting heterogeneous datasets into a common cell-embedding space. *Nature Communications*, 13, 2022.
59. X. Lin, S. Jiang, L. Gao, Z. Wei, and J. Wang. Multisc: A deep learning pipeline for analyzing multiomics single-cell data. *Briefings in Bioinformatics*, 25(6), 2024.
60. T. Ashuach, M. I. Gabitto, R. V. Koodli, G.-A. Saldi, M. I. Jordan, and N. Yosef. Multivi: Deep generative model for the integration of multimodal data. *Nature Methods*, 20:1222–1231, 2023.
61. L. Tarhan, J. Bistline, J. Chang, B. Galloway, E. Hanna, and E. Weitz. Single cell portal: An interactive home for single-cell genomics data. *bioRxiv*, 2023.
62. UCSC Genome Browser. Liftover tool. `https://genome.ucsc.edu/cgi-bin/hgLiftOver`, 2025.
63. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pages 8024–8035, 2019.
64. S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.
65. V. Svensson, K. N. Natarajan, L.-H. Ly, R. J. Miragaia, C. Labalette, I. C. Macaulay, A. Cvejic, and S. A. Teichmann. Power analysis of single-cell rna-sequencing experiments. *Nature Methods*, 14:381–387, 2017.
66. P. V. Kharchenko, L. Silberstein, and D. T. Scadden. Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11:740–742, 2014.
67. A. F. Agarap. Deep learning using rectified linear units (relu), 2018.
68. L. Prechelt. Early stopping – but when? In A. Krogh and et al., editors, *Neural networks: Tricks of the trade*, volume 7700 of *Lecture Notes in Computer Science*, pages 53–67. Springer, 2012.
69. D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.
70. J. Liu, Y. Huang, R. Singh, J.-P. Vert, and W. S. Noble. Jointly embedding multiple single-cell omics measurements. *Algorithms for Bioinformatics*, 143:10, 2019.
71. T. O. Kvålseth. On normalized mutual information: Measure derivations and properties. *Entropy*, 19(11):631, 2017.
72. J. M. Santos and M. Embrechts. On the use of the adjusted rand index as a metric for evaluating supervised classification. In *Artificial Neural Networks – ICANN 2009*, volume 5769 of *Lecture Notes in Computer Science*, pages 175–184. Springer, 2009.
73. 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.
74. A. Gayoso, Z. Steier, R. Lopez, J. Regier, K. L. Nazor, A. Streets, and N. Yosef. Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nature Methods*, 18:272–282, 2021.
75. T. Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.
76. J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science Engineering*, 9(3):90–95, 2007.
77. M. L. Waskom. Seaborn: Statistical data visualization, 2021. Version 0.11.2 or later, available at `https://seaborn.pydata.org/`.
78. Z. Fang, X. Liu, and G. Peltz. Gseapy: A comprehensive package for performing gene set enrichment analysis in python. *Bioinformatics*, 39(1):btac757, 2023.
79. M. Kanehisa, M. Furumichi, Y. Sato, M. Kawashima, and M. Ishiguro-Watanabe. Kegg for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research*, 51:D587–D592, 2023.
80. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.
81. R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33:495–502, 2015.
82. V. Svensson. Droplet scrna-seq is not zero-inflated. *Nature Biotechnology*, 38(2):147–150, 2020.

83. G. Eraslan, L. M. Simon, M. Mircea, N. S. Mueller, and F. J. Theis. Single-cell rna-seq denoising using a deep count autoencoder. *Nature Communications*, 10(1):390, 2019.
84. J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.
85. M. Setoguchi, N. Nasu, S. Yoshida, Y. Higuchi, S. Akizuki, and S. Yamamoto. Mouse and human cd14 (myeloid cell-specific leucine-rich glycoprotein) primary structure deduced from cdna clones. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, 1008(2):213–222, 1989.
86. N. Saito, K. A. Pulford, J. Breton-Gorius, J. M. Massé, D. Y. Mason, and E. M. Cramer. Ultrastructural localization of the cd68 macrophage-associated antigen in human blood neutrophils and monocytes. *The American Journal of Pathology*, 139(5):1053–1059, 1991.
87. V. Sisirak, N. Vey, B. Vanbervliet, T. Duhen, I. Puisieux, B. Homey, E. P. Bowman, G. Trinchieri, B. Dubois, D. Kaiserlian, S. A. Lira, A. Puisieux, J.-Y. Blay, C. Caux, and N. Bendriss-Vermare. Ccr6/ccr10-mediated plasmacytoid dendritic cell recruitment to inflamed epithelia after instruction in lymphoid tissues. *Blood*, 118(19):5130–5140, 2011.
88. Y.-C. B. Wu, D. Kipling, and D. K. Dunn-Walters. The relationship between cd27 negative and positive b cell populations in human peripheral blood. *Frontiers in Immunology*, 2:81, 2011.
89. W. Cao, S. H. Lee, and J. Lu. Cd83 is preformed inside monocytes, macrophages and dendritic cells, but it is only stably expressed on activated dendritic cells. *Biochemical Journal*, 385(1):85–93, 2005.
90. C. Bryant, P. D. Fromm, F. Kupresanin, G. Clark, K. Lee, C. Clarke, P. A. Silveira, H. Suen, R. Brown, E. Newman, and I. Cunningham. A cd2 high-expressing stress-resistant human plasmacytoid dendritic-cell subset. *Immunology Cell Biology*, 94(5):447–457, 2016.
91. S. F. Ziegler, F. Ramsdell, and M. R. Alderson. The activation antigen cd69. *Stem Cells*, 12(5):456–465, 1994.
92. T. H. Sulahian, P. Högger, A. E. Wahner, K. Wardwell, N. J. Goulding, C. Sorg, A. Droste, M. Stehling, P. K. Wallace, P. M. Morganelli, and P. M. Guyre. Human monocytes express cd163, which is upregulated by il-10 and identical to p155. *Cytokine*, 12:415–422, 2000.
93. E. P. Mimitou, A. Cheng, A. M. Montalbano, S. Hao, M. Stoeckius, M. Legut, T. Roush, A. Herrera, E. Papalexi, Z. Ouyang, R. Satija, and P. Smibert. Scalable and multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nature Biotechnology*, 37:1293–1301, 2019.
94. Satija Lab. Integrating scrna-seq and scatac-seq data using seurat, 2024. Seurat v5 Tutorial, Accessed: 2024-02-18.
95. Z. Miao, J. Wang, K. Park, D. Kuang, and J. Kim. Depth-corrected multi-factor dissection of chromatin accessibility for scatac-seq data with pacs. *Nature Communications*, 16:401, 2025.
96. Muon Team. Chromatin accessibility processing (pbmc 10k tutorial), 2024. Muon Tutorials, Accessed: 2024-02-19.
97. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. 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.
98. M. Kanehisa. Toward understanding the origin and evolution of cellular organisms. *Protein Science*, 28:1947–1951, 2019.
99. M. Kanehisa and S. Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28:27–30, 2000.
100. The Gene Ontology Consortium. The gene ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 2023.