PNAS RESEARCH ARTICLE | GENETICS
BIOPHYSICS AND COMPUTATIONAL BIOLOGY OPEN ACCESS

Robust probabilistic modeling for single-cell multimodal mosaic integration and imputation via scVAEIT

Jin-Hong Du[a], Zhanrui Cai[b], and Kathryn Roeder[a,c,1]

Contributed by Kathryn Roeder; received August 22, 2022; accepted November 3, 2022; reviewed by Wei Sun and Xiang Zhou

Recent advances in single-cell technologies enable joint profiling of multiple omics. These profiles can reveal the complex interplay of different regulatory layers in single cells; still, new challenges arise when integrating datasets with some features shared across experiments and others exclusive to a single source; combining information across these sources is called mosaic integration. The difficulties lie in imputing missing molecular layers to build a self-consistent atlas, finding a common latent space, and transferring learning to new data sources robustly. Existing mosaic integration approaches based on matrix factorization cannot efficiently adapt to nonlinear embeddings for the latent cell space and are not designed for accurate imputation of missing molecular layers. By contrast, we propose a probabilistic variational autoencoder model, scVAEIT, to integrate and impute multimodal datasets with mosaic measurements. A key advance is the use of a missing mask for learning the conditional distribution of unobserved modalities and features, which makes scVAEIT flexible to combine different panels of measurements from multimodal datasets accurately and in an end-to-end manner. Imputing the masked features serves as a supervised learning procedure while preventing overfitting by regularization. Focusing on gene expression, protein abundance, and chromatin accessibility, we validate that scVAEIT robustly imputes the missing modalities and features of cells biologically different from the training data. scVAEIT also adjusts for batch effects while maintaining the biological variation, which provides better latent representations for the integrated datasets. We demonstrate that scVAEIT significantly improves integration and imputation across unseen cell types, different technologies, and different tissues.

mosaic integration | multiomics | transfer learning | deep generative models

Significance

Single-cell multimodal assays provide an unprecedented opportunity for investigating heterogeneity of cell types and novel associations with disease and development. Although analyses of such multimodal datasets have the potential to provide new insights that cannot be inferred with a single modality, access typically requires the integration of multiple data sources. We propose a probabilistic variational autoencoder model for mosaic integration, which involves merging data sources that include features shared across datasets and features exclusive to a single data source. Our model is designed to provide a lower dimensional representation of the cells for visualization, clustering, and other downstream tasks; accurate imputation of missing features and observations; and transfer learning for robustly imputing new datasets when only partial measurements are available.

With new technological advances, researchers are able to measure a growing number of molecular dimensions, including the genome, transcriptome and epigenome, on millions of cells. The primary goals are to classify subtypes of cells, to understand cell function, and to model basic biological processes such as early development and clinically relevant traits, such as disorders and cancer. Integrating data from multiple modalities (1) and whole-genome measurements presents new challenges in the analysis of single-cell data. While no single technology can measure all relevant omics in a single cell, recent developments facilitate the measure of several; for example, TEA-seq (2) and the DOGMA-seq (3) simultaneously measure chromatin accessibility, gene expressions, and protein abundances. However, obtaining single-cell multimodal datasets that measure many modalities may be costly and this limits the sample sizes. Moreover, there is a need to integrate new data sources with existing multimodal atlases and impute the missing biological modalities. Therefore it is of fundamental importance to develop methodologies that can perform integrative analysis and cross-modal translation on the full range of jointly profiled multimodal single-cell datasets.

To integrate single-cell multimodal datasets, most existing methods identify anchors either explicitly or implicitly. Depending on the choice of anchor, the integration methods can be divided into three categories: horizontal integration, vertical integration, and diagonal integration (4). Horizontal integration methods, including Seurat v3's CCA (5), Harmony (6), and LIGER (7), use common modalities and features as anchors to link datasets containing different cells. Vertical integration approaches combine different modalities from datasets measured across a common set of cells. Representative works include 1) Seurat v4's Weighted Nearest Neighbor (WNN) (8), which identifies influential pairs of features based on the relative utility of each data modality and maps query datasets to the reference dataset based on shared variable features, and 2) totalVI (9), which models paired gene and protein measurements. On the other hand, there are methods that perform horizontal and vertical integration to combine different samples and modalities simultaneously. For example, MultiVI (10) models paired and unpaired

Author contributions: J.-H.D., Z.C., and K.R. designed research, performed research; J.-H.D. analyzed data; and J.-H.D., Z.C., and K.R. wrote the paper.

Reviewers: W.S., Fred Hutchinson Cancer Research Center; and X.Z., University of Michigan.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: roeder@andrew.cmu.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2214414119/-/DCSupplemental.

Published December 2, 2022.

measurements of gene and chromatin accessibility. Diagonal integration is considerably more challenging because neither modality nor cells are assumed to be shared. Extra cellular information is required to align different modalities.

As noted by Argelaguet et al. (4), a more general and challenging modern integration task is *mosaic integration*. For this integration task, different data modalities are profiled in different subsets of cells, or different subsets of cells are profiled from different experiments or technologies. Mosaic integration has two goals: 1) to map multimodal data to a common latent space, which is achieved by obtaining a joint multimodal profile for each cell, while utilizing the information from both shared and unshared features; and 2) to transfer knowledge of a fully trained model to a new dataset with partial modalities and features measured. This latter is sometimes also known as transfer learning (11, 12), and it is helpful in aligning cells and imputing unmeasured modalities and features for new datasets at a relatively low cost. Imputation of missing measurements and features is a natural bi-product of the integration procedure.

Recent advances in mosaic integration mainly focus on finding a common latent space while de-emphasizing the importance of imputation accuracy. For example, the non-negative matrix factorization algorithm UINMF (13) can align single-cell datasets containing both shared and unshared features in the low dimensional space; StabMap (14) first obtains low-dimensional score matrices for different datasets and then re-weights them based on features shared with the reference dataset. One common limitation of the current approaches is that only simple linear relationships between modalities are captured in the model (4). And it is hard to incorporate different types of covariates for batch effect adjustment while retaining biological variation (15–17). Besides, it is difficult for matrix factorization algorithms to align new datasets without training on the new dataset again. In the age of million single-cell multimodal data, both joint representations and transfer learning are indispensable. Although deep generative models such as totalVI (9) and MultiVI (10) can address the aforementioned limitations, they suffer from model misspecification issues when some features are missing for some cells. Specifically, existing deep generative methods either ignore unshared features or simply set the missing quantities to zero when performing integration, which induces biases. Failing to consider missing patterns, their performances will likely deteriorate when the proportion of unshared features increases significantly. As a result, new computational approaches that systematically model and utilize the information of missingness for combining mosaic-type multimodal datasets in the two scenarios are desirable.

We develop Single-Cell Variational AutoEncoder for Integration and Transfer learning (scVAEIT), a probabilistic deep learning algorithm (9, 18, 19) capable of performing mosaic integration and imputation. The model allows for arbitrary patterns of shared and unshared features and modalities, and the integration does not require the input of any extra biological information. In addition to great flexibility, a primary advantage of the approach is its robustness to overfitting. By incorporating a masking procedure, our model learns interpretable joint representations for cells and the distributions of unobserved features conditioned on an arbitrary subset of observed modalities and features. Unlike the traditional generative models that are fully unsupervised, imputing the masked features serves as a supervised learning procedure, which is analogous to the supervised PCA in Seurat's WNN (8). The masking procedure also acts as a form of regularization to mitigate overfitting, which is typical of the deep generative models. In contrast, conventional neural networks only focus on and remember, for example, genes and proteins that are easy to predict in multitask learning. scVAEIT is thus extremely useful for missing features imputation and crossmodal generation, providing great flexibility and high accuracy in learning a common latent space. Furthermore, it can robustly transfer crossmodal knowledge to new single-modal and multimodal datasets that only measure partial panels of features of the training datasets, providing generalization on transfer learning.

Results

Method Overview. An overview of the multimodal single-cell data analysis pipeline with scVAEIT is shown in Fig. 1. Existing single-cell data studies provide researchers with a variety of multimodal and single-modal datasets, illustrated here with three datasets of PBMCs (peripheral blood mononuclear cells) (3): DOGMA-seq (RNA, protein, peaks), CITE-seq (RNA, protein), and ASAP-seq (RNA, peaks). Building on deep generative models, scVAEIT provides a flexible way to jointly analyze multiple multimodal and single-modal single-cell datasets while incorporating additional covariates for batch effect adjustments. After the model is trained on the mosaic-type dataset, it can then be applied to various downstream tasks. Specifically, it enables joint latent representations (intermediate integration) of all modalities and transfer learning to new data sources (late integration). Most importantly, it can robustly impute the missing quantities of the mosaic-type datasets. Though the model we illustrate in this setting is applicable for general multi-omic settings.

The critical innovation of scVAEIT lies in the introduction of a mask M, incorporating information about missingness and the complementary observed features X_{M^c} (Fig. 1*B*). Instead of isolating shared and unshared features in multiple datasets, we utilize a missing mask M for each cell to inform scVAEIT about the missing pattern when performing mosaic integration. Hence, scVAEIT enables integrative analysis of cells from different sources with more or fewer features and modalities measured. It differs from other deep generative models in that it explicitly learns the conditional distributions of certain masked features (or modalities) given unmasked features (or modalities). In contrast, other variational inference models simply set the missing quantities as zero, which biases the learned models.

Even though we do not impose any assumption on the relationships between the shared and unshared features, scVAEIT learns the interdependence among features and modalities during the optimization process. This is achieved by sampling random mask M to force the model to predict the masked features based on the unmasked features. For example, masking out the peaks for cells in the DOGMA-seq datasets helps to impute the chromatin accessibility for cells in the CITE-seq datasets. This is also valuable for dealing with structural missing problems as we can exploit prior knowledge about data. On the other hand, randomly masking out a small portion of features also acts as a way of regularization, which prevents the model from overfitting the training dataset.

Because scVAEIT is optimized by the mini-batch stochastic gradient descent algorithm, its memory usage does not depend on the number of cells, which makes it scalable for large datasets. For instance, it can process the DOGMA-seq dataset with 13,763 cells and over 29,139 features (genes, proteins, and chromatin accessibility) within 1 h for intermediate integration on a single

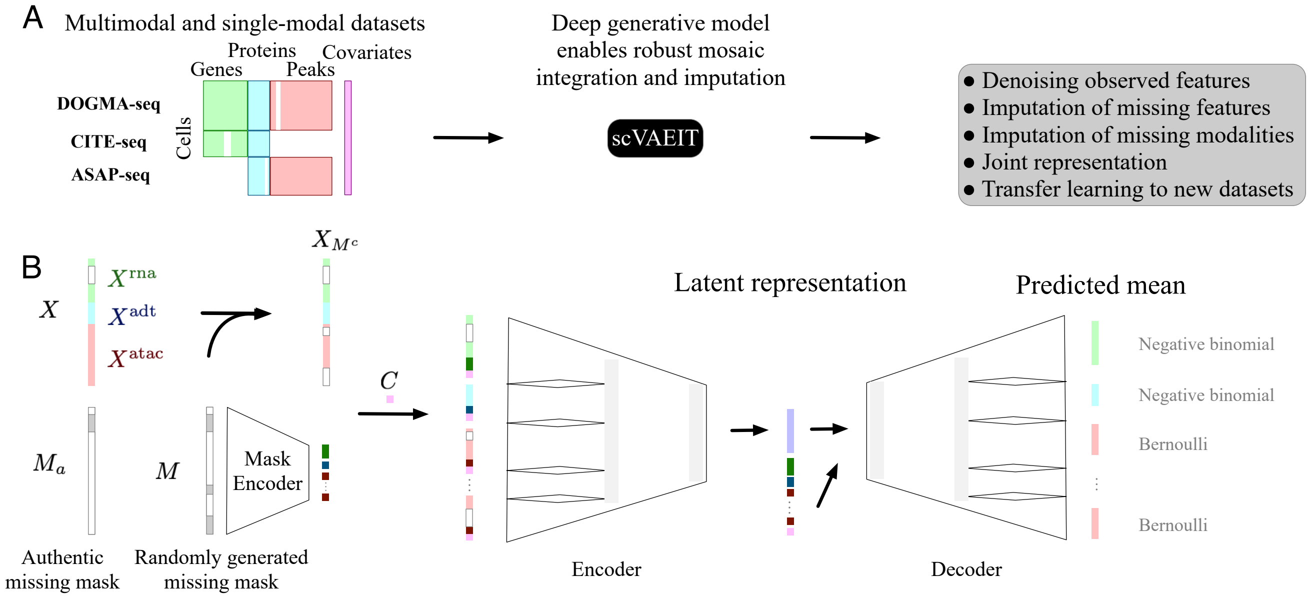


Fig. 1. Overview of multimodal single-cell datasets mosaic integration with scVAEIT. (*A*) Multimodal sequencing techniques such as DOGMA-seq, CITE-seq, and ASAP-seq simultaneously measure multiple modalities in single cells. Different datasets can measure different panels of features for the same modality, producing mosaic count matrices for RNA, proteins and peaks. scVAEIT takes these matrices and an optional matrix containing continuous and categorical covariates to learn cross-modality and cross-feature relationships. It also produces a joint representation of all modalities for each cell. (*B*) For each cell, scVAEIT uses a mask to inform the missingness of the data. The actual or authentic mask is M_a. During training, masks (M) are randomly generated and encoded for each modality to force the model to learn to predict specific portions of the features based on other observed features (X_{M^C}). In the first layer of the encoder and the last layer of the decoder, different modalities and peaks in different chromosomes break into subconnections that greatly reduces model complexity. The encoder then outputs the parameters of the posterior distributions for the latent variable Z, given the unmasked data X_{M^C} and the mask M. Next, samples from the posterior distributions along with the mask and covariates are fed to a decoder neural network to predict the posterior mean of X. Ultimately unobserved values are imputed and observed values are denoised.

Tesla v100-32 GPU. Once the model learns the cross-modality and cross-feature relationship from the training dataset, it can then robustly transfer its knowledge (late integration) to new sources at a relatively low cost. For example, denoising and imputing the previous DOGMA-seq dataset takes less than one minute on a single GPU. As more and more multimodal single-cell atlases become available, it is valuable to train scVAEIT on a reference atlas for once and readily transfer learns the new sources for cross-modality translation and imputation. Details of the specifications of the model architecture and training procedures are included in the *Method* section and *SI Appendix*.

Cross-Domain Translation with High Accuracy. To examine cross-domain translation accuracy, we used a dataset consisting of PBMCs processed by CITE-seq (8) (see the *Methods* section). We held out one cell type for evaluation, trained different models based on the remaining cells, and then imputed each modality given the other for each held-out cell type. The two largest cell types—Mono ($n = 49{,}010$ cells) and CD4 T ($n = 41{,}001$ cells)—were examined, and the results are summarized in Fig. 2*A*. As the held-out cell types were not present in the training set, high accuracy on cross-domain translation indicates that the model learns cross-modality relationships rather than memorizing the training set.

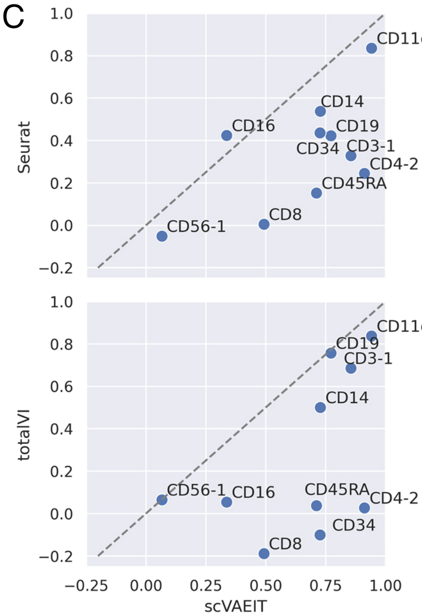
We compared the imputation accuracy with Seurat v4′s WNN (8) with multimodal anchor transfer and totalVI (9), a deep generative model designed for CITE-seq datasets (see the *Methods* section). In almost all cases, scVAEIT achieved higher correlation and lower RMSE compared to Seurat's WNN method and totalVI. For protein imputation, Seurat v4 WNN failed to extrapolate well on the unseen cell type because it is based on protein-gene similarity in the training set instead of learning the nonlinear relationship between proteins and genes as scVAEIT. On the other hand, totalVI performed worst on predicting protein counts given gene counts, as it is designed for joint representation of proteins and genes and naturally fails to translate between the two modalities. Overall, the results indicate that scVAEIT captures the cross-modality relationships very well and is accurate for imputing missing modalities.

Transfer Learning External Datasets. We next applied scVAEIT to impute proteins of two external datasets without any additional fine-tuning or training. Unlike the CITE-seq PBMC dataset we used to train the model, the CITE-seq CBMC (cord blood mononuclear cell) dataset was from a different tissue, while the REAP-seq PBMC dataset was generated from a different experimental protocol. Both of the two datasets measure genes and proteins as well. Such datasets are challenging because of the experimental biases and variances. Because both of the external datasets consist of a few proteins, it is unlikely any model can successfully impute gene expressions based on proteins. Therefore, we only inspected how well the models can infer the proteins in the external datasets when a partial panel of the genes is observed. On these external datasets, scVAEIT aligned more accurately with measured protein levels than Seurat and totalVI (Fig. 2*C* and *SI Appendix*, Fig. S1). More specifically, scVAEIT achieved strong positive correlation (median Pearson correlation = 0.73, median Spearman correlation = 0.69) and small RMSE of 1.70 Drop) on the CITE-seq CBMC dataset. At the same time, it was more stable than totalVI on imputation accuracy (Fig. 2*B*). We note that scVAEIT was not enforced to focus on only a few proteins in these external datasets, as it was trained on a panel of 227 proteins. The superior performance across tissues and technical protocols is due to the training scheme of random masking, which helps scVAEIT to give all-around attention to individual proteins.

A

Holdout Cell Type	Source	Target	Metric	Method: scVAEIT	Method: Seurat	Method: totalVI
Mono	RNA	ADT	r_p	**0.88**	0.79	0.53
			r_s	**0.85**	0.76	0.48
			RMSE	**0.72**	0.98	2.21
	ADT	RNA	r_p	**0.81**	**0.81**	0.66
			r_s	0.55	**0.56**	0.48
			RMSE	**0.47**	0.48	0.68
CD4 T	RNA	ADT	r_p	**0.82**	0.77	0.34
			r_s	**0.83**	0.76	0.31
			RMSE	**0.78**	0.98	2.29
	ADT	RNA	r_p	**0.80**	0.76	0.62
			r_s	**0.45**	0.43	0.40
			RMSE	**0.45**	0.49	0.70

C



B

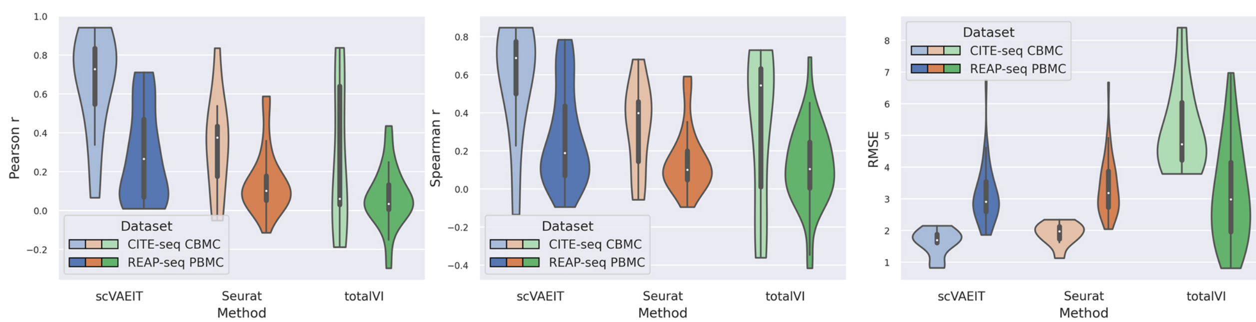


Fig. 2. Evaluation of protein imputation on two external datasets. (*A*) Performance of missing modality imputation on the hold-out cell types when the source modality is fully observed. The evaluation metrics include Pearson correlation (r_p), Spearman correlation (r_s) and root mean square error (RMSE). RNA and ADT (antibody derived tag) represent the gene and protein modalities, respectively. (*B*) Violin plots of the Pearson correlation, Spearman correlation and root mean square error (RMSE) between the imputed and the true protein abundances on CITE-seq CBMC dataset and REAP-seq PBMC dataset. (*C*) Across tissue correlations between imputed and measured protein abundance on the CITE-seq CBMC data: scVAEIT versus Seurat and scVAEIT versus totalVI.

Trimodal Integration and Imputation. The recently proposed DOGMA-seq protocol revealed distinct changes in different modalities during native hematopoietic differentiation and peripheral blood mononuclear cell stimulation. However, most deep generative models are restricted to bimodal analysis. For example, totalVI (9) jointly models genes and proteins, and MultiVI (1) jointly models genes and chromatin accessibility. Although the existing method BABEL (20) uses separate encoders and decoders to model each modality and could be extended to trimodal analysis, it is still conceptually hard to align different modalities in the latent space. The difficulty of multimodal analysis lies in combining and balancing information from different modalities; thus, we would like to examine how scVAEIT performs in such cases.

To quantify how leveraging trimodal information improves dataset imputation, we compared our method with totalVI and MultiVI. A stimulation indicator was provided for all three methods as an extra covariate for batch effect correction. We also included three-way weighted nearest neighbors (3-WNN) in Seurat v4 (8) as a benchmarking method, using Harmony (6) to integrate the stimulated and control data. We held out each cell type in the DOGMA-seq dataset as the test set and trained all models based on the remaining cells. Then the models were evaluated by imputing each modality given the other modalities (Fig. 3 and *SI Appendix*, Fig. S2).

scVAEIT achieved strong performance across different held-out cell types and experimental conditions. For gene imputation, other deep generative models could not extrapolate well on the unseen cell types, producing large RMSEs, while Seurat's anchor transfer method failed to capture the relationships between different modalities, resulting in low Pearson correlations and Spearman correlations (Fig. 3*A*). scVAEIT, however, excelled in both aspects for imputing the gene counts. The binary versus continuous nature of the three modalities required us to use different metrics for evaluating chromatin accessibility predictions. Inferring peaks from gene and protein expressions, it is on par with MultiVI on the area under the receiver operating characteristic (AUROC) while significantly better on accuracy (ACC) and RMSE (Fig. 3*B*). If we zoom in to look at the density scatterplot of the normalized protein counts and the imputed counts (Fig. 3*C*), we see that Seurat overestimated the protein abundances of CD8a in the held-out CD4 T cell type because the test cell type is similar to the CD8 T cell type, where the CD8a protein highly expresses. For imputation of sparse counts of genes MAML2, LINC00681, SOS1, FHIT, and MYBL1, we also observed Seurat's underestimation and over smoothing (*SI Appendix*, Fig. S3). These results indicate scenarios where Seurat's map and query method might fail. On the contrary, scVAEIT learned a nonlinear and more accurate mapping from the expressions of all available modalities and

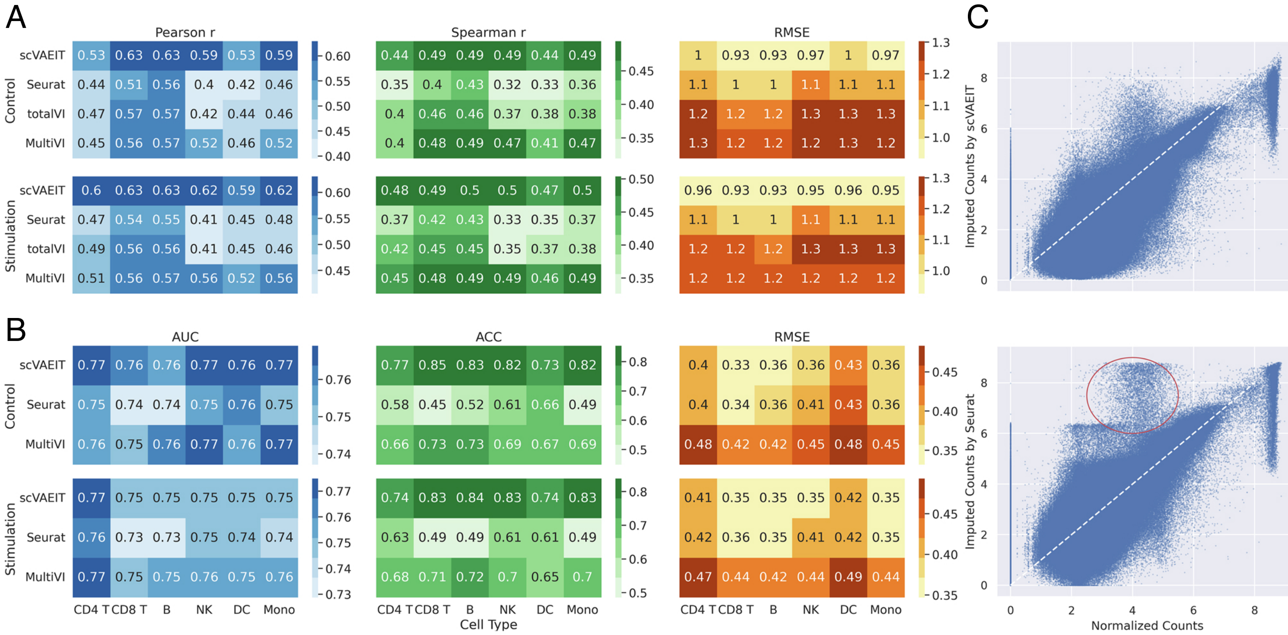


Fig. 3. Trimodal imputation analysis on DOGMA-seq dataset. (*A*) Performance of gene imputation on the held-out cell types of the DOGMA-seq dataset. (*B*) Performance of chromatin accessibility imputation on the held-out cell types of the DOGMA-seq dataset. (The metrics AUROC and ACC denote the area under the receiver operating characteristic and accuracy, respectively). (*C*) Scatter plot of normalized protein counts versus imputed values on the held-out CD4 T cell type of the DOGMA-seq dataset by scVAEIT and Seurat. Each dot represents one protein abundance for one CD4 T cell. Because CD4 T cells in the query dataset are most similar to CD8 T cells in the reference dataset, Seurat simply memorizes the expression patterns of CD8a protein (within the red circle) from CD8 T cells and consequently overestimates its expression for CD4 T cells.

features to the unknown features and hence achieved a more reasonable imputation on the missing quantities even in the unseen CD4 T cell type.

Robustness to Missing Features. As more and more single-cell multimodal atlases become available, we could expect that late integration would be increasingly essential and practical for researchers to transfer knowledge from the atlases to the new datasets; however, the new datasets may not contain all the features measured in a reference atlas, leading to missing data issues in practice. Therefore, we investigated how different levels of missingness of the measured features might impact the late integration of new datasets by using the DOGMA-seq dataset. In the largest CD4 T cluster, a specific portion of genes, proteins, and chromatin accessibility, was randomly held out as a test set, and different models were trained on cells in other cell types and then applied to impute these missing quantities based on the observed features in the held-out cell type. The process was repeated multiple times.

Except for scVAEIT, we observed that other deep generative models have unstable behaviors on imputation accuracy under the model misspecified scenario (Fig. 4 *A* and *B* for proteins and chromatin accessibility imputation, and *SI Appendix*, Fig. S4 for genes imputation). More specifically, the performance of totalVI on gene and protein imputation deteriorated dramatically when the missing proportion increased; MultiVI, on the other hand, had a larger variance and uncertainty when facing a larger degree of missingness. On the other hand, as a nonparametric method, Seurat was more stable with respect to different levels of missingness (Fig. 4*A*). It obtained better Pearson correlations for gene imputation but worse Spearman correlation than MultiVI (*SI Appendix*, Fig. S4), meaning that Seurat did not capture nonlinear relationships well among sparse signals. As noted, scVAEIT effectively combined the advantages of both methods, and was robust and accurate even when features were missing and the training model was misspecified to the new datasets.

Application to Multi-Source and Multimodal Mosaic Integration. We further applied scVAEIT to integrate the DOGMA-seq dataset with a CITE-seq PBMC dataset and an ASAP-seq dataset from Mimitou et al. (3). After filtering low-quality cells and features (see the *Methods* section), the three datasets have 208 proteins in common, while the DOGMA-seq and the CITE-seq datasets have only 880 shared genes, and the DOGMA-seq and the ASAP-seq datasets have only 26,206 shared peaks (*SI Appendix*, Table S1). We first considered the task of two-phase mosaic integration, where the mosaic multimodal datasets were combined through intermediate integration to remove the effects of experiment conditions, and the new mosaic multimodal datasets were imputed afterwards (late integration). Each cell type from the three datasets was held out for imputation and evaluation in turns, while the rest were used to perform intermediate integration. As all datasets measure a shared panel of proteins, it is easier to use protein counts to link these datasets together. Instead, we inspected how protein counts can be imputed based on other modalities in the held-out cell type.

We compared scVAEIT with Seurat's 3-WNN and anchor transfer method. Although the procedure of two-phase integration is straightforward for scVAEIT, it becomes much more complicated for Seurat's method. First, the multimodal reference and query datasets need to be integrated with Harmony to adjust for the stimulation effects separately, when performing dimension reduction. Then the query dataset is mapped to the reference dataset in the low-dimensional space. Finally, the missing quantities are imputed based on the similarity between the reference and query cells computed in the low-dimensional space. In our experiments, the multimodal neighbor method failed for integrating cells in the DC or Mono cell type because

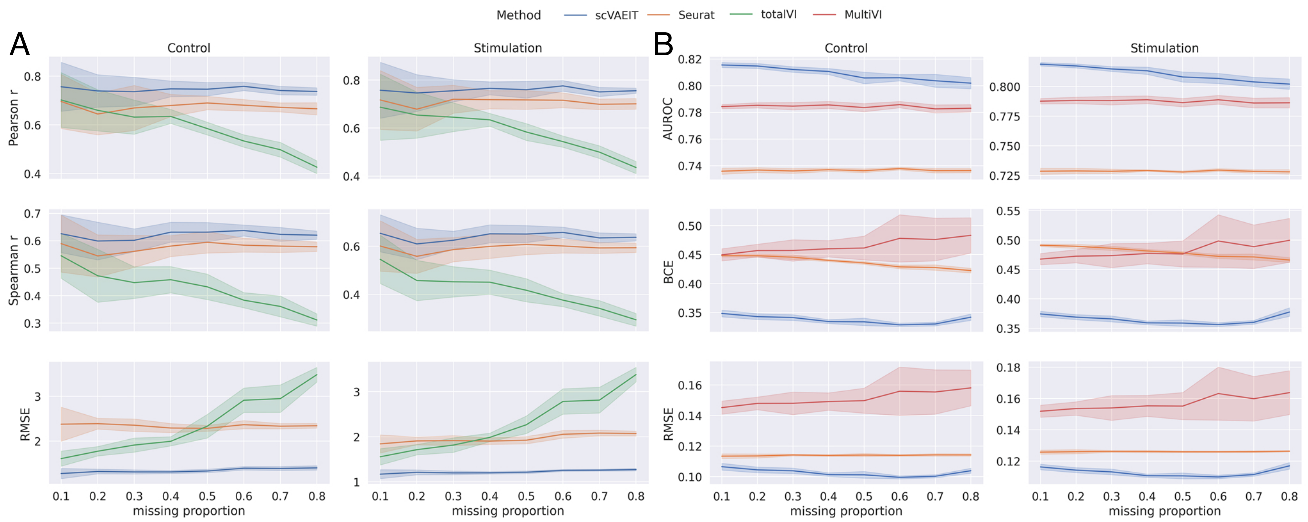


Fig. 4. Incorporating masking information during training enables robust late integration with missing features. (*A*) Performance of protein imputation on the held-out CD4 T cell type of the DOGMA-seq dataset with random missing. For each missing proportion, all methods are evaluated using the same observed features across 10 runs and the shaded region is within one SD of the average performance. (*B*) Performance of chromatin accessibility imputation on the held-out CD4 T cell type of the DOGMA-seq dataset with random missing (the metrics AUROC and BCE denote the area under the receiver operating characteristic and binary cross entropy, respectively). For each missing proportion, all methods are evaluated using the same observed features across 10 runs and the shaded region is within one SD of the average performance.

there were too few cells, no matter how we chose the number of neighbors. As shown in Fig. 5*A*, scVAEIT achieved consistently better performance on different cell types in terms of Pearson correlation, Spearman correlation, and RMSE. The overall better performance across the three different multimodal datasets also indicated that scVAEIT learns to infer protein counts based on gene expressions (CITE-seq), chromatin accessibility (ASAP-seq), or both (DOGMA-seq).

Next, we refitted both models using all cell types and visualized the cell embeddings in Fig. 5*B*. We also compared with matrix the factorization method UINMF (13) (*SI Appendix*, Fig. S5*A*). We performed Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) (21) directly on the learned latent variables for scVAEIT, while a similar UMAP visualization based on the trimodal WNN graph was also shown for Seurat. *SI Appendix*, Fig. S6*A* shows the same embeddings colored by annotated cell types obtained from the original paper (3). The majority of the CD4 T and the CD8 T cells, as identified in the original paper, reside in the upper right half and the lower right half of each cluster in Fig. 5*B*, respectively. Because a portion of the cells in all datasets was stimulated with anti-CD3/CD28, the stimulation-dependent changes within T cells can be observed. More specifically, the activated CD4 T and CD8 T cells clusters (clusters 5 and 16 in figure 5G of the original paper (3)) can be easily recognized in scVAEIT's embeddings. On the other hand, integration based on Harmony and Seurat's WNN removed not only specific batch effects but also meaningful biological signals: notably, it is not easy to identify stimulated T cells from Seurat's embeddings; UINMF failed to adjust for batch effects correctly. By contrast, scVAEIT retained meaningful biological differences when adjusting for the effect of experimental conditions. When we visualized the source of datasets, we observed that scVAEIT merges cells from different datasets more evenly than Seurat. In scVAEIT's embeddings, the cells from different datasets were aligned based on the expression of features; hence, they were not entirely mixed because of the differences in expression distribution from different techniques. If the researchers assure that such an effect should be eliminated, one can further apply batch effect correction methods such as FastMNN (22) on the latent variables and map all cells to the same source.

The dynamic changes in CD3, CD279, and CD69 protein abundances of control and stimulated cells were also examined in Fig. 5*C* and *SI Appendix*, Figs. S5*B* and S6 *B* and *C*. CD3 protein is highly expressed (green) in the CD4 T cell type (the top right cluster) and the CD8 T cell type (the bottom right cluster) in the absence of stimulation, while the differences between cell types become smaller in the stimulated cells, as expected. In the major CD4 T cell type, we noticed that the expression level of control cells gradually varied in scVAEIT's embeddings while it remained almost the same in Seurat's embeddings. Similar results can also be observed for CD279 and CD69 proteins (*SI Appendix*, Fig. S6 *B* and *C*), which have different expression patterns in the control and the stimulated cells. Furthermore, scVAEIT also provided a convenient way to visualize both denoised and imputed gene expressions, especially when a gene is only measured in some of the datasets (*SI Appendix*, Fig. S7). For example, the CD3E gene is only available in the CITE-seq dataset (*SI Appendix*, Fig. S7*A*), while the CD69 gene is only available in the DOGMA-seq and CITE-seq datasets (*SI Appendix*, Fig. S7*B*). The denoised and imputed expressions can then be used to test and identify differential features on the integrated dataset (*SI Appendix*, Fig. S10). Furthermore, this procedure can be naturally generalized for differential expression testing between case and control groups by applying the Bayesian inference framework (23, 24).

Discussion

scVAEIT is a highly adaptable procedure, designed for multiple integration tasks, including joint representations (intermediate integration 1), and transfer learning and imputation on new datasets (late integration 1). First, scVAEIT utilizes information from shared and unshared features of different modalities, such as genes, proteins, and chromatin accessibility, to build an integrative probabilistic model for intermediate integration. The proposed model learns complex nonlinear relationships between modalities, enables a common latent representation of different

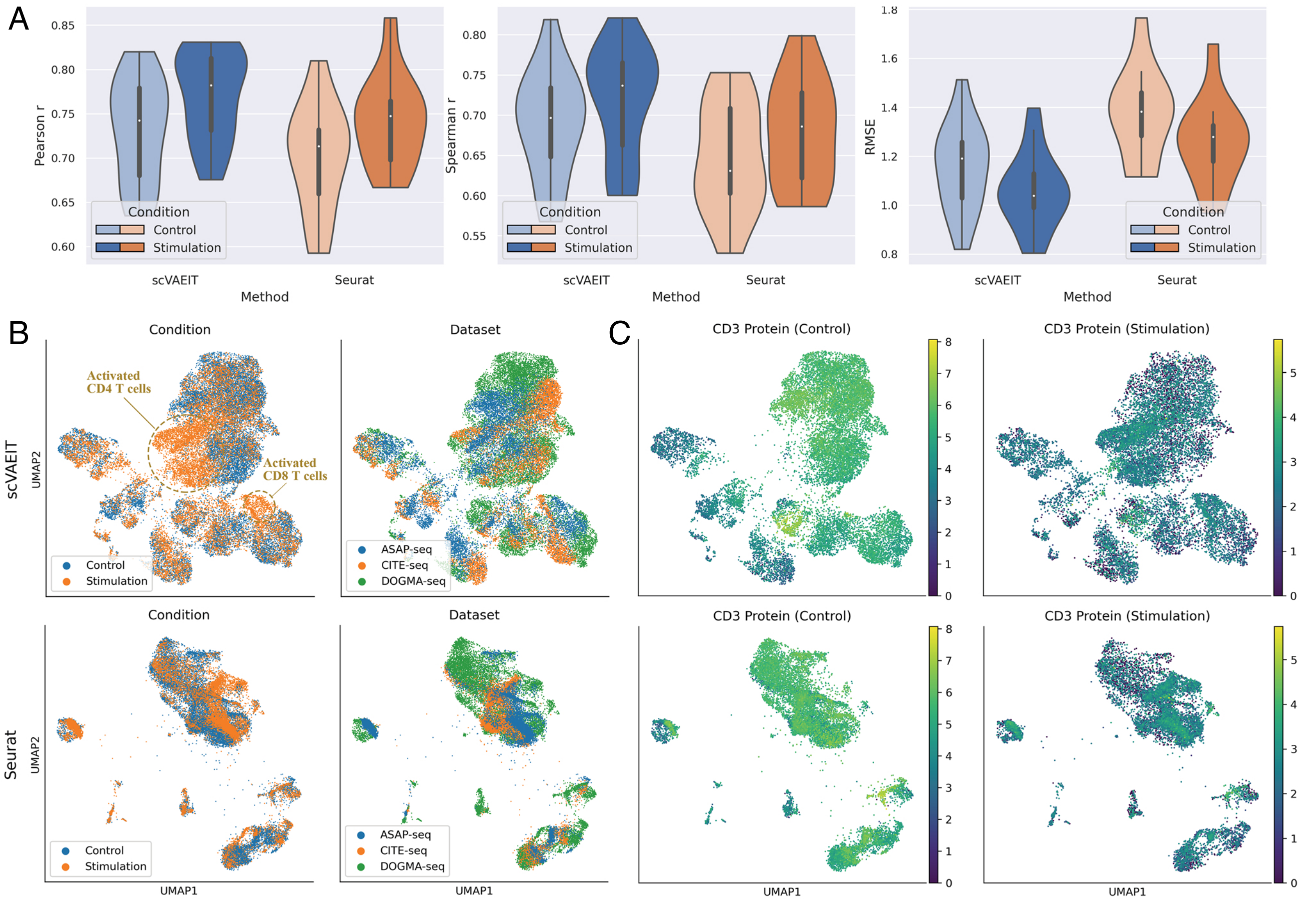


Fig. 5. Integration of DOGMA-seq, CITE-seq, and ASAP-seq PBMC datasets. (*A*) Performance of protein imputation after two-phase integration with held-out cell types of three multimodal datasets. The two smallest cell types of the three datasets - DC and Mono - contain too few cells for Seurat's integration method to find multimodal nearest neighbors, even with a reduced number of multimodal neighbors. (*B*) Joint embeddings of the three multimodal datasets after intermediate integration. Seurat's WNN with Harmony overcorrects the experimental condition effect, eliminating the difference between the stimulated and control T cells. (*C*) Log-normalized expressions of CD3 protein in the control cells (first column) and the stimulated cells (second column) on scVAEIT's and Seurat's embeddings. Purple and yellow dots correspond to lowly and highly expressed cells, respectively.

modalities for each cell, and can accurately impute the missing quantities. Second, after training on some data source, scVAEIT readily transfers its knowledge to a new data source, even when the new dataset only contains partial measurements of features or modalities of the training data, enabling robust cross-modality translation. Third, scVAEIT is flexible in incorporating different types of covariates and adjusting for batch effects when combining single-cell datasets from various experiments, tissues, and technologies. Surprisingly, even including single-modal datasets can help with multimodal learning for scVAEIT, as revealed in *SI Appendix*, Fig. S9. Compared to other deep generative models, the success of scVAEIT relies on the masking procedure that helps the model predict a certain portion of features in a supervised manner. This not only makes great usage of mosaic-type datasets but also serves a role of regularization.

The success of deep generative models (9, 10, 20) on single-cell multimodal data analysis enables expressive and scalable probabilistic representations of cells. Such models can harness the strengths of each modality by combining multiple cellular views in an end-to-end pipeline. The deep neural networks are adequate to describe complex and heterogeneous relationships between cells and modalities, and the probabilistic modeling provides interpretability with uncertainty quantification. Despite these promising results, the deep generative models can suffer from model misspecification issues when facing mosaic datasets. This happens when a model is used to perform intermediate integration on multiple multimodal datasets with different missing measurements or during late integration when a learned model is applied to transfer knowledge to new datasets that measure different panels of features from the training dataset. In experimental results, when the degree of model misspecification increases, the accuracy and stability of other deep generative models can be hugely impacted, and their performance can be inferior to nearest neighbor methods if the information of missingness is not taken into consideration.

During the training process of scVAEIT, the masking procedure plays a vital role in learning conditional distributions under arbitrary missing patterns. The supervised nature and regularization effect of the masking procedure produce more robust model constructions when learning from mosaic datasets and transferring knowledge to new datasets. There is still flexibility in choosing masks for training. For example, we can incorporate specific structural missing patterns to generate the masks with prior information on the predictable relationship between features. It will be valuable when we try to efficiently and systematically integrate more and more modalities.

Materials and Methods

Datasets Preprocessing. *SI Appendix*, Table S1 summarizes the information of the preprocessed datasets. For the CITE-seq PBMC dataset (8), we first identified the top 5,000 variable genes and then filtered out genes and proteins expressed in less than 500 cells. The filtered dataset consists of 161,764 cells with 4,686 genes and 227 proteins. For the CITE-seq CBMC dataset (25), we first filtered cells that have less than 200 genes and then filtered out genes that are expressed in less than 10 cells. The filtered dataset consists of 7,891 cells with 3,464 genes and 10 proteins that are shared in the CITE-seq PBMC dataset. For the REAP-seq PBMC dataset (26), we first filtered cells with less than 200 genes or more than or equal to 5% mitochondrial counts and then filtered out genes expressed in less than 500 cells. The filtered dataset consists of 7,092 cells with 3,864 genes and 38 proteins shared in the CITE-seq PBMC dataset. For genes and proteins in the DOGMA-seq PBMC dataset, we applied the same procedure as above, resulting in 2,166 genes and 208 proteins. For chromatin accessibility, we retained the 75% variable peaks and filtered out peaks that appear in less than 500 cells or are on sex chromosomes, leaving 26,765 peaks. The CITE-seq PBMC and the ASAP-seq PBMC datasets from Mimitou et al. (3) are filtered analogously.

After the low-quality cells and features were filtered, we size-normalized the gene and protein counts separately, such that each cell's counts sum to 10,000 counts per cell. Then we log-transformed the size-normalized counts. For peaks, we binarized them by replacing all nonzero values with a value of 1. Preprocessing external single-cell expression datasets with different measurements for transfer learning and model evaluation required computing the median size factor for observed measurements from the training set and performing log-normalization afterward.

Probabilistic Modeling of Multimodal Datasets. Inspired by recent advancement on conditional variational inference (18, 27, 28) in the machine learning community, we aim to model the missing features and missing modalities problem altogether as a conditional probability estimation problem. Consider m modalities measured in the single cells. For each cell, we denote its measurement by $X = (X^1, \ldots, X^m) \in \mathbb{R}^d$ where $X^i \in \mathbb{R}^{d_i}$ are samples of ith modality with $d = d_1 + \cdots + d_m$. We introduce a binary mask $M \in \{0, 1\}^d$ for X and its bitwise complement M^c, such that the jth entry of the observed sample X_{M^c} is X_j if $M_j = 1$ and 0 otherwise. Then, X_M is defined to be $X - X_{M^c}$. The authentic missing pattern M_a represents which components of X are actually missing, while the distribution of M can be arbitrary during training. For example, if we want to model missing completely at random, the entries of M could be independent Bernoulli random variable. Furthermore, we can incorporate extra structural information to model the situation of missing modality. To model the conditional distribution of the observed modalities given the missing values or modalities, we consider the following maximum likelihood problem:

$$\max_{\theta} \mathbb{E}_{X,M} \log p_\theta(X_M \mid X_{M^c}, M).$$

In other words, we aim to determine the conditional distribution of X_M given X_{M^c} and M. Since there are totally 2^d missing patterns, we can learn 2^d conditional distributions of $X_M|X_{M^c}, M$ separately, each of them is a function: $X_{M^c} \mapsto X_M$ for a given M. However, this is computationally infeasible. Instead, we jointly model all conditional distributions by using a single neural network. Since neural networks are universal approximators (28, 29), a neural network can well approximate arbitrarily complex function: $(X_{M^c}, M) \mapsto X_M$ when it has enough capacity and non-polynomial activation functions. Therefore, we expect that a single neural probabilistic model can approximate *all* conditional distributions of unobserved features conditioned on any subset of observed features.

Since the above condition density itself is hard to formulate and optimize, we follow the variational Bayesian approach (30) to maximize the negative evidence lower bound (ELBO) instead:

$$\log p_\theta(X_M \mid X_{M^c}, M) \geq \underbrace{\mathbb{E}_{q_\psi(Z|X,M)} \log p_{\theta_2}(X_M \mid Z, X_{M^c}, M)}_{\mathcal{L}_{\mathrm{impute}}}$$
$$-KL(q_\psi(Z \mid X, M) \| p_{\theta_1}(Z \mid X_{M^c}, M)) := \mathcal{L}_M, \qquad [\mathbf{1}]$$

where $Z \in \mathbb{R}^m$ is a latent variable with approximate posterior distribution q_ψ, KL denotes the Kullback–Leibler divergence, and $\theta = (\theta_1, \theta_2)$. We specify the distributions for data as follows.

In this paper, we consider trimodal analysis that includes gene expressions, protein abundances and chromatin accessibility, though the method is readily applied for more general settings. The gene counts for the nth cell are represented by a G-dimensional vector $X_n^{\mathrm{rna}} = (X_{ng}^{\mathrm{rna}})_{g\in[G]}$ such that X_{ng}^{rna} is the observed RNA count of gene g in cell n. Likewise, an A-dimensional protein counts vector $X_n^{\mathrm{adt}} = (X_{ng}^{\mathrm{adt}})_{a\in[A]}$ denotes the observed protein counts and an P-dimensional binary vector $X_n^{\mathrm{atac}} = (X_{np}^{\mathrm{atac}})_{p\in[P]}$ representing the occurrence of peaks for cell n. Let $Z_n \in \mathbb{R}^m$ and M_n be the associated joint latent variable and mask for cell n.

Under the target distribution p_{θ_1}, we assume that the latent variables are normally distributed:

$$Z_n \mid X_{nM_n^c}, M_n \sim \mathcal{N}(\mu_{\theta_1}, \mathrm{diag}(\sigma^2_{\theta_1,1}, \ldots, \sigma^2_{\theta_1,m})). \qquad [\mathbf{2}]$$

Ideally, we want Z_n generated from p_{θ_1} to be as close as possible to the one generated from the the proposal distribution q_ψ when X_n is fully observed except for its authentic missing entries:

$$Z_n \mid X_{nM_a^c}, M_a \sim \mathcal{N}(\mu_\psi, \mathrm{diag}(\sigma^2_{\psi,1}, \ldots, \sigma^2_{\psi,m})). \qquad [\mathbf{3}]$$

This formulation also allows us to compute the KL divergence analytically in the ELBO Eq. **1**, while it is possible to extend to normal mixtures to model more complex latent structures (19). In our implementation, we simply set $q_\psi(Z \mid X_{M_a^c}, M_a) = p_{\theta_1}(Z \mid X_{M^c}, M)$ to reduce computational complexity. Finally, q_ψ and p_{θ_2} are modeled as two fully-factorized Gaussian distributions, whose mean and variance are estimated by two neural networks respectively. The generative distribution p_{θ_1} are also assumed to be fully-factorized for X_M given Z, X_{M^c} and M. We use negative binomial (NB) distribution to model the gene expression and the protein abundance. We assume that the counts are generated based on Z_n as follow

$$X_{ng}^{\mathrm{rna}} \mid Z_n, M_n \sim \mathrm{NB}(\lambda_{ng}^{\mathrm{rna}}, \theta_g^{\mathrm{rna}}), \qquad [\mathbf{4}]$$
$$X_{na}^{\mathrm{adt}} \mid Z_n, M_n \sim \mathrm{NB}(\lambda_{na}^{\mathrm{adt}}, \theta_a^{\mathrm{adt}}), \qquad [\mathbf{5}]$$
$$X_{np}^{\mathrm{atac}} \mid Z_n, M_n \sim \mathrm{Bernoulli}(\mu_{np}^{\mathrm{atac}}), \qquad [\mathbf{6}]$$

which are independent of M given Z_n. Here the parameters $\lambda_{ng}^{\mathrm{rna}}$ and θ_g^{rna} are the expected total count and the inverse dispersion of the negative binomial distribution, and the parameters $\lambda_{ng}^{\mathrm{adt}}$ and θ_g^{adt} are defined analogously for protein counts. For each peak, μ_{np}^{atac} represents its posterior mean. The posterior expectations $\lambda_{ng}^{\mathrm{rna}}$, $\lambda_{na}^{\mathrm{adt}}$, and μ_{np}^{atac} are outputted by the decoder, while the dispersion parameters are treated as trainable variables. These parameters are learned from the data.

The aforementioned probabilistic modeling Eq. **1** emphasizes missing features and modalities imputation. On the other hand, we not only want to impute the unobserved quantities, but also want to denoise the observed quantities. Therefore, we also attempt to maximize the reconstruction likelihood

$$\mathcal{L}_{rec} := \mathbb{E}_{p_{\theta_2}(Z \mid X_{M^c}, M)} \log p_{\theta_1}(X_{M^c} \mid Z, M). \qquad [\mathbf{7}]$$

Network Architecture. scVAEIT is implemented using the Tensorflow (31) (version 2.4.1) Python library. scVAEIT consists of three main branches, the mask encoder, the main encoder and the main decoder. For each cell n, a missing mask M_n is embeded as E_n to a short dense vector through the mask encoder, which greatly reduces the input dimension to the main encoder and decoder. Then, the encoder takes data X_n (log-normalized gene and protein counts, and binary peaks), a mask embedding vector E_n and (optional) covariates C_n as input, and output the estimated posterior mean and variance of the distribution of the latent variable Z_n. Next, a realization is draw from this posterior distribution and fed to the decoder along with the mask embedding vector E_n and the covariates C_n. The decoder finally outputs the posterior mean of X.

We use subconnections at the first layer of the encoder and the last layer of the decoder. In each of these layer, there are 256 and 128 units for genes and proteins respectively, and 16 units for peaks in each chromosome. The weights are isolated between different blocks. The mask encoder outputs 32-dim, 16-dim, and 2-dim vectors for genes, proteins, and peaks in each chromosome. Besides these special layers, we also have one fully-connected hidden layer of 256 units in the encoder and the decoder, with LeakyReLU activation functions.

Model Training. scVAEIT is trained in an end-to-end manner. The objective function is a convex combination of the ELBO Eq. **1** and the reconstruction likelihood Eq. **7**:

$$\boldsymbol{\mathcal{L}} := \beta \boldsymbol{\mathcal{L}}_M + (1-\beta)\boldsymbol{\mathcal{L}}_{recon},$$

where $\beta \in [0, 1]$ is a hyperparameter set to be 0.5 for all experiments. That is, the parameters are optimized by Monte Carlo sampling to maximize the weighted average of the reconstruction likelihood and the imputation likelihood, while minimizing the KL divergence between masked posterior latent variable $Z \mid X_{M^c}, M$ and the authentic posterior latent variable $Z \mid X_{M_a^c}, M_a$. During training, with equal probability we observe the original data and the masked data. The mask is repeatedly randomly generated for each cell at the beginning of every gradient update step in each epoch during the optimization process, such that each modality is observed with equal probability, and each entry is further randomly masked out with probability 0.2. The sensitivity analysis on the masking probability is also included in *SI Appendix*, Fig. S8. To balance the magnitudes of different modalities, we calculated the weighted likelihood using weights $(w^{\mathrm{rna}}, w^{\mathrm{adt}}) = (0.15, 0.85)$ for bimodal datasets and $(w^{\mathrm{rna}}, w^{\mathrm{adt}}, w^{\mathrm{atac}}) = (0.14, 0.85, 0.01)$. The default variable initializer in Tensorflow is used, sampling weight matrix from a uniform distribution and setting bias vectors to be zero. We train our model for 300 epochs using the AdamW optimizer (32), a variant of the stochastic gradient descent algorithm, with a batch size of 512, a learning rate of 10^{-3}, and a weight decay of 10^{-4}. We also use batch normalization to aid in training stability. Because we use mini-batches for training, scVAEIT's memory usage does not effected by the number of cells. Instead, it is only related to in the number of features in the dataset and number of neural network parameters.

Benchmarking Methods. We compare scVAEIT with Seurat (8), totalVI (9), and MultiVI (10). Seurat v4's WNN is used to perform intermediate integration and multimodal anchor-based transfer-learning method is used to perform transfer learning to new datasets. Standard preprocessing procedures in Seurat are used for evaluating Seurat's results. More specifically, RNA counts are normalized by LogNormalization method and protein counts are normalized by centered log-ratio (CLR) method. When evaluating Seurat's protein imputation result, we revert the CLR normalized imputed counts and perform log-normalization. The log-normalized gene counts and size-normalized protein counts are provided as input to totalVI; the log-normalized gene counts and binary peaks are provided as input to MultiVI. For stimulation effect correction, we use Harmony (6)'s corrected dimension reductions for running Seurat's WNN, and provide an indicator variable as a covariate to totalVI and MultiVI. For running Harmony and 3-WNN integration on DOGMA-seq datasets, we use the script provided in the original paper (3) (https://github.com/caleblareau/asap_reproducibility/blob/master/pbmc_stim_multiome/code/11_setup.R). For running UINMF (13) (version 1.1.0) on the three multimodal datasets, we first split each dataset into two based on stimulation conditions. Then each of the six datasets are preprocessed with functions `normalize` and `scaleNotCenter` according to their tutorials. After that, the shared features and unshared features of the six datasets are separated and supplied to function `optimizeALS`, where the parameters are chosen by examining the results of functions `suggestK` and `suggestLambda`. The imputed values are obtained with function `imputeKNN`. For running totalVI and MultiVI, we set the latent dimension as 32, `early_stopping_patience` as 15 and leaving all other hyperparameters as default. All code required to reproduce our reported results, including data preprocessing and model training, have been deposited on GitHub (https://github.com/jaydu1/scVAEIT).

Evaluation Metric. We use multiple evaluation metrics for comparing different methods on imputing gene counts, protein counts and peaks. As the log-normalized gene and protein expressions are treated as continuous, we use Pearson correlation and Spearman correlation to evaluate their imputation quality. For n observations $(x_1, \ldots, x_n)$ and their prediction/imputation values $(y_1, \ldots, y_n)$, the correlation metrics are defined as

$$r_p = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}},$$

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)},$$

where $\bar{x}$ and $\bar{y}$ are the mean of x_i's and y_i's respectively, and $d_i = \mathrm{rank}(x_i) - \mathrm{rank}(y_i)$ is the difference between the two ranks of each observation. The imputation error is also quantified by the RMSE,

$$RMSE(x, y) = \sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - y_i)^2}.$$

For chromatin accessibility, the peaks are binarized while the imputed values are continuous in $[0, 1]$. Thus we use area under the receiver operating characteristic (AUROC), binary cross entropy (BCE) and RMSE to evaluate imputation accuracy of binary peaks. The AUROC metric takes value between 0 and 1, which is commonly used in statistics and machine learning community. A larger value of AUROC indicates that the binary outcome is easier to predict based on imputed value at various threshold settings. The BCE metric

$$BCE(x, y) = -\frac{1}{n}\sum_{i=1}^n (x_i \log y_i + (1 - x_i)\log(1 - y_i)),$$

is equivalent to the negative log-likelihood of Bernoulli variables. Thus a smaller value of BCE means a better fit of statistical models. For each evaluation metric, we effectively consider each gene, protein, or peak in each cell a separate observation.

Data, Materials, and Software Availability. All datasets used in this paper are previously published and freely available. For bimodal datasets integration, CITE-seq PBMC cells from Hao et al. (8) are in the GEO database under accession code GSE164378 and a Seurat object containing filtered cells is also provided in their tutorial https://satijalab.org/seurat/articles/multimodal_reference_mapping.html; the CITE-seq CBMC cells from Stoeckius et al. (25) are in the GEO database under accession code GSE100866 and a Seurat object containing these cells is also available as 'bmcite' in the SeuratData (v0.2.1) package; the REAP-seq PBMC cells from Peterson et al. (26) are in the GEO database under accession code GSE100501. For trimodel datasets integration, the DOGMA-seq, CITE-seq, and ASAP-seq PBMC cells from Mimitou et al. (3) are in the GEO database under accession code GSE156478, and the intermediate result files are retrieved from their Github repository https://github.com/caleblareau/asap_reproducibility. The Python package of scVAEIT is publicly available at https://github.com/jaydu1/scVAEIT with MIT license. Python and R scripts for reproducing all results in this paper are also provided in the same repository.

ACKNOWLEDGMENTS. We thank the referees for helpful suggestions. This work used Bridges-2 system at the Pittsburgh Supercomputing Center (PSC) through allocations MTH210011 and BIO220140 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which of the former is supported by NSF grants OAC-1928147. This project was funded by National Institute of Mental Health (NIMH) grant R01MH123184.

Author affiliations: [a]Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213; [b]Department of Statistics, Iowa State University, Ames, IA 50011; and [c]Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA 15213

1. Z. Miao, B. D. Humphreys, A. P. McMahon, J. Kim, Multi-omics integration in the age of million single-cell data. *Nat. Rev. Nephrol.* **17**, 710–724 (2021).
2. E. Swanson *et al*., Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *eLife* **10** (2021).
3. E. P. Mimitou *et al*., Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat. Biotechnol.* **39**, 1246–1258 (2021).
4. R. Argelaguet, A. S. Cuomo, O. Stegle, J. C. Marioni, Computational principles and challenges in single-cell data integration. *Nat. Biotechnol.* **39**, 1202–1215 (2021).
5. T. Stuart *et al*., Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
6. I. Korsunsky *et al*., Fast, sensitive and accurate integration of single-cell data with harmony. *Nat. Methods* **16**, 1289–1296 (2019).
7. T. Stuart, R. Satija, Integrative single-cell analysis. *Nat. Rev. Genet.* **20**, 257–272 (2019).
8. Y. Hao *et al*., Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
9. A. Gayoso *et al*., Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nat. Methods* **18**, 272–282 (2021).
10. T. Ashuach, M. I. Gabitto, M. I. Jordan, N. Yosef, Multivi: Deep generative model for the integration of multi-modal data. bioRxiv (2021).
11. J. Wang *et al*., Data denoising with transfer learning in single-cell transcriptomics. *Nat. Methods* **16**, 875–878 (2019).
12. Z. Zhou, C. Ye, J. Wang, R. Zhang, Surface protein imputation from single cell transcriptomes by deep neural networks. *Nat. Commun.* **11**, 1–10 (2020).
13. A. R. Kriebel, J. D. Welch, UINMF performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization. *Nat. Commun.* **13**, 1–17 (2022).
14. S. Ghazanfar, C. Guibentif, J. C. Marioni, Stabmap: Mosaic single cell data integration using non-overlapping features. bioRxiv (2022).
15. H. T. N. Tran *et al*., A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21**, 1–32 (2020).
16. M. D. Luecken *et al*., Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
17. S. K. Chu, S. Zhao, Y. Shyr, Q. Liu, Comprehensive evaluation of noise reduction methods for single-cell RNA sequencing data. *Brief. Bioinform.* **23**, bbab565 (2022).
18. D. P. Kingma, M. Welling, "Auto-encoding variational Bayes" in *2nd International Conference on Learning Representations*, Y. Bengio, Y. LeCun, Eds. (2014).
19. J. H. Du, M. Gao, J. Wang, Model-based trajectory inference for single-cell RNA sequencing using deep learning with a mixture prior. bioRxiv (2020).
20. K. E. Wu, K. E. Yost, H. Y. Chang, J. Zou, Babel enables cross-modality translation between multiomic profiles at single-cell resolution. *Proc. Natl. Acad. Sci. U.S.A.* **118** (2021).
21. L. McInnes, J. Healy, N. Saul, L. Großberger, Umap: Uniform manifold approximation and projection. *J. Open Source Soft.* **3**, 861 (2018).
22. L. Haghverdi, A. T. Lun, M. D. Morgan, J. C. Marioni, Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
23. R. Lopez, P. Boyeau, N. Yosef, M. Jordan, J. Regier, Decision-making with auto-encoding variational Bayes. *Adv. Neural Inform. Proc. Syst.* **33**, 5081–5092 (2020).
24. P. Boyeau *et al*., An empirical Bayes method for differential expression analysis of single cells with deep generative models. bioRxiv (2022).
25. M. Stoeckius *et al*., Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
26. V. M. Peterson *et al*., Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.* **35**, 936–939 (2017).
27. K. Sohn, H. Lee, X. Yan. "Learning Structured Output Representation using Deep Conditional Generative Models." *NIPS* (2015).
28. O. Ivanov, M. Figurnov, D. Vetrov, "Variational autoencoder with arbitrary conditioning" in *International Conference on Learning Representations* (2018).
29. K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators. *Neural Netw.* **2**, 359–366 (1989).
30. D. M. Blei, A. Kucukelbir, J. D. McAuliffe, Variational inference: A review for statisticians. *J. Am. Stat. Assoc.* **112**, 859–877 (2017).
31. M. Abadi *et al*., TensorFlow: Large-scale machine learning on heterogeneous systems (2015). Software available from tensorflow.org.
32. I. Loshchilov, F. Hutter, "Decoupled weight decay regularization" in *International Conference on Learning Representations* (2017).