The performance of deep generative models for learning joint embeddings of single-cell multi-omics data

Eva Brombacher [1,2,3,4,5,†] Maren Hackenberg [1,2,†] Clemens Kreutz [1,2,4]
Harald Binder [1,2] Martin Treppner [1,2,*]

[1]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, Freiburg, Germany
[2]Freiburg Center for Data Analysis and Modeling, University of Freiburg, Freiburg, Germany
[3]Spemann Graduate School of Biology and Medicine (SGBM), University of Freiburg, Freiburg, Germany
[4]Centre for Integrative Biological Signaling Studies (CIBSS), University of Freiburg, Freiburg, Germany
[5]Faculty of Biology, University of Freiburg, Freiburg, Germany

† These authors have contributed equally to this work and share first authorship.
* Corresponding author: treppner@imbi.uni-freiburg.de

June 6, 2022

Abstract

Recent extensions of single-cell studies to multiple data modalities raise new questions regarding experimental design. For example, the challenge of sparsity in single-omics data might be partly resolved by compensating for missing information across modalities. In particular, deep learning approaches, such as deep generative models (DGMs), can potentially uncover complex patterns via a joint embedding. Yet, this also raises the question of sample size requirements for identifying such patterns from single-cell multi-omics data. Here, we empirically examine the quality of DGM-based integrations for varying sample sizes. We first review the literature and give a short overview of deep learning methods for multi-omics integration. Next, we consider six popular tools in more detail and examine their robustness to different cell numbers, covering two of the most common multi-omics types currently favored. Specifically, we use data featuring simultaneous gene expression measurements at the RNA level and protein abundance measurements for cell surface proteins (CITE-seq), as well as data where chromatin accessibility and RNA expression are measured in thousands of cells (10x Multiome). We examine the ability of the methods to learn joint embeddings based on biological and technical metrics. Finally, we provide recommendations for the design of multi-omics experiments and discuss potential future developments.

Keywords: multi-omics, deep learning, experimental design, sample size, transcriptomics, proteomics, epigenomics

1 Introduction

Many diseases, such as cancer, affect complex molecular pathways across different biological layers. Consequently, there is currently an ongoing surge in multi-omics techniques that study the interaction of biomolecules across various omics layers [42, 54]. Multi-omics techniques have been used, e.g., to infer mechanistic insights about molecular regulation, the discovery of new cell types, and the delineation of cellular differentiation trajectories [1, 11, 49, 53]. However, because performing multi-omics experiments in the same cell is still costly and experimentally complex, many experiments have been carried out with comparatively small numbers of cells so far. Additionally, single-cell multi-omics data suffer from the sparseness and noisiness of the measured modalities, differences in sequencing depth, and batch effects. Data analysis is further complicated by differing feature spaces as well as shared and modality- or batch-specific variation [30].

Deep learning approaches, known for their ability to learn complex non-linear patterns from data, have become a popular building block for integrating different data types [15, 21]. For example, in 2021's Conference on Neural Information Processing Systems (NeurIPS) competition (`https://openproblems.bio/neurips_2021`), which addressed the topic of multimodal single-cell data integration, neural networks proved to be the most popular, with shallow deep learning models being among the best-performing methods [30]. Specifically, deep generative models (DGMs) such as variational autoencoders (VAEs) are increasingly employed to infer joint embeddings, i.e., low-dimensional representations, from multi-omics datasets, allowing all further downstream analyses to be performed simultaneously within this joint latent space (**Figure 1 B**).

This review provides a systematic overview of current DGM-based approaches for learning joint embeddings from multi-omics data and illustrates how small sample sizes impact what can be recovered from multi-omics datasets. Specifically, we examine how the performance of popular DGM-based approaches to infer joint low-dimensional representations from such data is influenced by varying numbers of cells. The required number of cells is particularly relevant at the stage of designing an experiment [51]. To tackle the challenging task of evaluating the quality of a latent representation with respect to the conservation of biological signal and batch correction capabilities, we draw on the guidelines provided by Luecken et al. [35].

The training of DGMs on multi-omics data is challenging due to the inherent high dimensionality and low sample size of multi-omics data and the large number of model parameters that need to be estimated while avoiding overfitting and bias [23]. Thus, we investigate the impact of cell numbers on the performance of selected single-cell multi-omics integration algorithms. We consider six popular VAE-based tools that incorporate different integration paradigms and training strategies for this illustration. Specifically, we included product-of-experts- and mixture-of-experts-based approaches and techniques that employ additional, commonly used integration techniques: cross-modality translation and adversarial training. Also, we chose models with different degrees of architectural complexity, including one model [31] with (self-)attention modules and additional regularization by clustering consistency. We, thus, created an exemplary selection that represents the range of architectural choices, additional training and regularization strategies, and levels of complexity currently used for the task at hand. Thus, viewing the selected models as representatives of the current landscape of DGMs for multi-omics integration, our case study allows us to draw conclusions from the performance of the

investigated tools in small sample size scenarios, and to give recommendations regarding architectural choices, integration strategies, and regularization paradigms.

2 Deep learning background

In recent years, deep learning has emerged as a potent tool for analyzing high-throughput biological data. At the core of such approaches are artificial neural networks (ANNs) that provide powerful yet versatile building blocks to learn complex non-linear transformations and thus uncover underlying structures from high-dimensional data. While supervised deep learning relies on labeled data to solve, e.g., classification problems, unsupervised deep learning can be employed in exploratory analyses to uncover central structure in data. For example, researchers frequently aim to understand cell-type compositions, for which they usually rely on unlabelled data. Hence, unsupervised deep learning methods have become increasingly popular in omics data analysis. Specifically, DGMs have been used for imputation [33, 62], visualization of the underlying structure of single-cell RNA-sequencing (scRNA-seq) data [14], and synthetic data generation [39, 51].

DGMs are trained to capture the joint probability distribution over all features in the input data, thus allowing to also generate new synthetic data with the same patterns as the training data by sampling from the learned distribution. This is typically done by introducing latent random variables z in addition to the observed data x. In single-cell transcriptomics applications, these latent variables might encode complex gene programs based on non-linear relationships between genes. Typically, the joint distribution $p_\theta(x, z)$ of observed and latent variables is described through a parametric model, where θ represents the model parameters. The joint probability can be factorized into a prior probability $p_\theta(z)$ and a posterior $p_\theta(x|z)$ and can thus be written as $p_\theta(x, z) = p_\theta(z)p_\theta(x|z)$. Inferring the data likelihood $p_\theta(x) = \int p_\theta(x, z)dz$ from the joint distribution requires marginalizing over all possible values of z, which is typically computationally intractable [27]. Hence, approximate inference techniques are employed to efficiently optimize the model parameters [7].

In single-cell applications, the most frequently used DGMs to date are Variational autoencoders (VAEs) [26] and generative adversarial networks (GANs) [20], which we present in more detail below.

2.1 Variational autoencoders

VAEs employ two independently parameterized but jointly optimized neural network models to learn an explicit parametrization of the underlying probability distributions. This is achieved by non-linearly encoding the data into a lower-dimensional latent space and reconstructing back to the data space. Specifically, the encoder (or recognition model) maps the input data x to a lower-dimensional representation given by a sample of the latent variable z, while the decoder network performs a reverse transformation and aims to reconstruct the input data based on the lower-dimensional latent representation (**Figure 1 A**).

To approximate the underlying data distribution $p_\theta(x)$, the encoder and decoder parameterize the conditional distributions $p_\theta(z|x)$ and $p_\theta(x|z)$, respectively. Since $p_\theta(x)$ and $p_\theta(z|x)$ are intractable, a variational approximation $q_\phi(z|x)$ is employed, typically given by a Gaussian distribution with diagonal covariance matrix.

Intuitively, the model is trained by reconstructing its inputs based on the lower-dimensional data representation, such that the latent space recovers the central factors of variation that allow for approximating the data distribution as closely as possible. Formally, a training objective for the model can be derived based on variational inference [7]. The parameters ϕ and θ of the encoder and decoder distributions can be optimized by maximizing the evidence lower bound (ELBO), a lower bound for the true data likelihood $p_\theta(x)$, with respect to ϕ and θ. Denoting with KL the Kullback-Leibler divergence $\mathrm{KL}\left[q\|p\right] := \mathbb{E}_q[\log \frac{q}{p}]$ for probability distributions q and p, the ELBO is given by

$$ELBO(x;\phi,\theta) = \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] = \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - \mathrm{KL}\left[q_\phi(z|x)||p(x)\right] \quad (1)$$

Here, the likelihood of a single observation (i.e., cell) x indicates how well it is supported by the model.

The decoder network is typically built to learn the parameters of specific distributions, which best describe the underlying biological data. For scRNA-seq and surface protein data (CITE-seq) a negative binomial distribution is frequently assumed, while single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq) data usually requires an additional modeling term that accounts for the increased sparsity of the data, e.g., in the form of a zero-inflated negative binomial (ZINB) distribution [40]. Other approaches use a binarized version of the scATAC-seq data [3, 57, 63, 68].

The typical workflow for analyzing high-dimensional (single- or multi-) omics data with a VAE is illustrated in Figure 1 B. The data is embedded with the encoder to obtain a low-dimension representation, which can subsequently be used for downstream analysis, such as clustering or trajectory inference.

2.2 Multimodal variational autoencoders

Several approaches already exist in which multimodal VAEs Shi et al. [45] are used to map different omics measurements into a common latent representation [19, 34, 40]. Each of these methods uses different approaches to combine the latent variables of the respective modalities. We can usually distinguish between mixture of experts (MoE) and product of experts (PoE) models (**Figure 1 C**). Hence, we describe a MoE and a PoE model in more detail below and examine their performance in our analyses.

We denote a single-cell multimodal dataset as $x_{1:M}$, where two modalities ($M = 2$) is the most common case. The joint generative model can therefore be written as $p_\theta(x_{1:M}, z) = p(z)\prod_{m=1}^{M} p_{\theta_m}(x_m|z)$, where $p_{\theta_m}(x_m|z)$ represents the likelihood of the decoder network for modality m, and $\theta = \{\theta_1, \ldots, \theta_M\}$.

For the MoE model, the resulting joint variational posterior can be factorized into $q_\phi(z|x_{1:M}) = \sum_{m=1}^{M} \alpha_m q_{\phi_m}(z|x_m)$, with $\alpha_m = 1/M$ and $\phi = \{\phi_1, \ldots, \phi_M\}$. This results in the following ELBO:

$$\begin{aligned} ELBO &= \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{z_m \sim q_{\phi_m}(z|x_m)}\left[\log \frac{p_\theta(x_{1:M}, z)}{q_\phi(z|x_{1:M})}\right] \\ &= \frac{1}{M}\sum_{m=1}^{M}\left\{\mathbb{E}_{z_m \sim q_{\phi_m}(z|x_m)}\left[\log p_\theta(x_{1:M}|z_m)\right] - KL\left[q_{\phi_m}(z|x_m)||p(x)\right]\right\} \end{aligned} \quad (2)$$

In contrast, PoE approaches [19, 34] combine the variational posteriors of the individual modalities

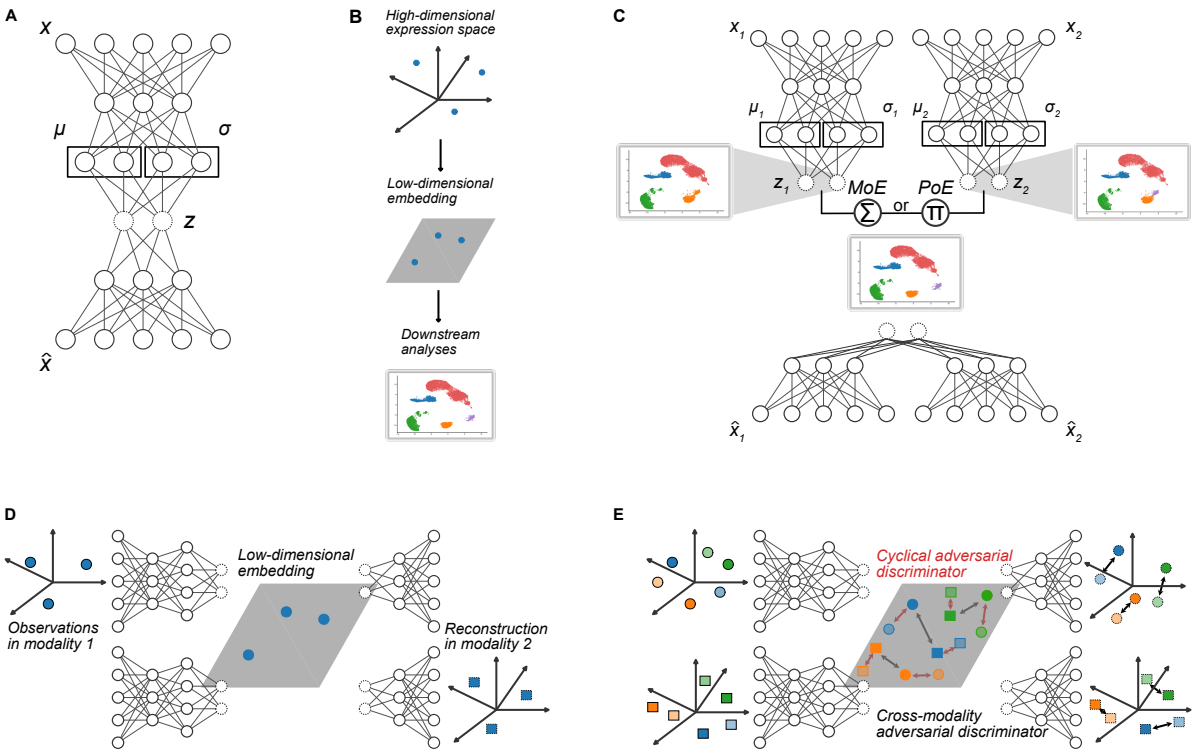

Figure 1: **Neural network architectures.** (A) Exemplary network architecture of vanilla VAE, where x represents the input data and $\hat{x}$ is the reconstructed data. Random variables z in the bottleneck layer are indicated by dashed circles. μ and σ represent the mean and standard deviation of the distributions, typically Gaussian distributions with diagonal covariance matrices, learned in the bottleneck layer. (B) Typical workflow: High-dimensional omics data are mapped to a low-dimensional embedding, which can then be utilized for visualization and downstream analyses such as clustering or trajectory inference. (C) General architecture of multimodal VAEs. (D) Cross-modality translation: High-dimensional measurements from one modality are mapped to a low-dimensional embedding with the modality-specific encoder. The latent representation is then used as input for the decoder of the respective other modality. (E) Adversarial training principles: Adversarial discriminators can be employed (1) to align low-dimensional embeddings of different modalities (squares vs. circles) of the same cell (same color) in the latent space (black arrows), (2) to align reconstructed profiles with the cross-modal reconstructions (lighter colors) obtained by decoding low-dimensional embeddings of one modality with the decoder of the other modality (black arrows in the reconstruction space, or (3) to align re-embedded decoder outputs from intra-modal and/or cross-modal reconstruction (lighter colors) with the original embeddings (red arrows in the latent space).

as products $q_\phi(z|x_{1:M}) = \prod_{m=1}^{M} q_{\phi_m}(z|x_m)$.

Shi et al. [45] argue that PoE approaches suffer from potentially overconfident experts, i.e., experts with lower standard deviations will tend to have a more considerable influence on the combined posterior, as experts with lower precision come with lower marginal posteriors. In contrast, in the MoE approach we consider here, both modalities receive equal weighting, reflecting the assumption that both modalities are of similar importance. Intuitively, employing a PoE approach corresponds to taking the 'intersection' of the individual posteriors, as a single posterior assigning a near-zero likelihood to a specific observation is enough to cause the product to be near-zero. In contrast, an MoE approach corresponds to taking the 'union' of all posteriors. Additionally, the weights α_m assigned to each modality can be adjusted to

reflect prior assumptions on their relative importance or be learned from the data during training.

2.3 Cross-modality translation

In addition to architectural choices regarding the integration of the modality-specific sub-networks via a PoE or MoE approach, many VAE-based methods introduce training objectives that facilitate specific functionality such as cross-modality translation or encourage particular properties of the embedding, such as clustering consistency between the modality-specific latent representations. On a higher level, these components can be seen as regularizers that push the embeddings found by the model towards certain desired properties.

A prominent example of such an additional feature to direct a joint embedding is cross-modality translation. Here, a cell's measurements of one modality, say, gene expression, are mapped to the joint latent space with the respective modality-specific encoder. Then, the decoder of another modality, say, chromatin accessibility, is employed to map the latent representation of the gene expression profile to a corresponding chromatin accessibility profile (Figure 1 D). This is only possible due to the integration of both modalities into a shared latent space, in which one cell's encoded representations of different modalities align.

When paired measurements of both modalities in the same cell are available, the translated reconstructions in the respective other modality can be compared to the cell's observed profile during training. The model learns a latent embedding that facilitates consistent cross-modality predictions. Thus, the model is explicitly pushed towards an embedding from which both modality-specific profiles can be reconstructed equally well, and that can, therefore, help in better capturing general underlying biological cell states as defined by the interplay of both modalities.

After training, cross-modality translation can be used to impute measurements of cells for which a specific modality is missing or to answer counterfactual questions such as "based on this specific gene expression profile, what would the corresponding chromatin accessibility profile have looked like?". This could be further combined with in-silico perturbations, i.e., generating synthetic profiles of one modality and using the model to infer corresponding profiles in other modalities. Additionally, this technique can be used to query the trained model for, e.g., subpopulations of cells where the cross-modality predictions are particularly well or particularly poorly aligned with the true measurements and to further characterize them, thus, also facilitating interpretability.

Examples of approaches that employ this technique are given by, e.g., Wu et al. [57], Minoura et al. [40] and Zhao et al. [65], and will be presented in more detail below in Section 3 and in the experimental Section 6

2.4 Adversarial training strategies

Another commonly used regularization technique is given by adversarial training, which is closely related to cross-modality translation and is often employed concurrently. Such adversarial components are often integrated into a variational or standard autoencoder framework. Inspired by generative adversarial networks (GANs) [20], a discriminator, typically specified as a neural network, is employed to distinguish between two omics modalities, either based on the samples from their latent representations or based on reconstructed samples from cross-modal decoders (Figure 1 E, black arrows). The objective

of the discriminator then is to maximize the probability of correctly identifying the original modality a samples comes from, while the encoder and decoder of the (V)AE model are trained to fool the discriminator by producing samples that are indistinguishable to the discriminator. Thus, the (V)AE model is encouraged to find a latent embedding in which the different modalities are better aligned and integrated, and/or learn decoders that allow for accurate cross-modal predictions, that are well aligned with the intra-modal predictions. This is achieved by incorporating additional adversarial penalty terms into the loss function and thus training the discriminators jointly with the (V)AE model.

Additionally, such adversarial components can be used to train the model in a cyclical fashion for additional intra-modal and cross-modal consistency. For intra-modal consistency, the low-dimensional embeddings of samples of one modality are decoded with the modality-specific decoder. Subsequently, the reconstructions are re-encoded with the modality-specific encoder and compared to the original embedding of the sample. An adversarial discriminator can be employed to align the embedding of the original sample with the embedding of the re-encoded reconstruction of that sample for intra-modal consistency of the latent space (Figure 1 E, red arrows). For cross-modal consistency, the low-dimensional embeddings from samples of one modality are decoded with the decoder of the other modality and subsequently re-encoded with the encoder of the other modality. By aligning these cross-modal embeddings with the original embeddings using an adversarial discriminator, the model can learn to produce cross-modal translations that are consistent with the original sample when re-embedded in the latent space.

More generally, such regularization by adversarial components is based by the training strategy of GANs [20]. Just as VAEs, GANs are another form of DGMs that differ from VAEs predominantly in how the joint probability distribution over all input features is specified. While VAEs learn an explicit parameterization of (an approximation to) this distribution (see 2.1), in GANs, this distribution is available only implicitly via sampling. A GAN consists of a generator and a discriminator neural network that can be thought of as playing a zero-sum minimax game: The generator simulates synthetic observations that are presented to the discriminator together with real data observations. The discriminator then has to decide whether a given sample is a real observation or a synthetic one from the generator. The two networks are trained jointly, with the discriminator trying to maximize the probability of correctly distinguishing between synthetic and real observations, and the the generator trying to fool the discriminator, i.e., to generate samples that are indistinguishable from the real ones.

In multi-omics data integration, such adversarial approaches are typically integrated into (V)AE models as additional components to regularize the latent representation and/or the decoder reconstructions [22, 60, 65], while a purely GAN-based approach is given by, e.g., Amodio and Krishnaswamy [2].

3 Literature review

Although recently, several available deep learning-based applications for the integration of single-cell multi-omics data have been reviewed in [15] and [46], there is still a lack of a more comprehensive review focusing specifically on DGMs. In the following, we are going to survey approaches for paired (both modalities measured in the same cell in one experiments) and unpaired (modalities measured in different cells in separate experiments) single-cell data. An overview is given in Table 1, where we

list recent deep learning-based approaches for multi-omics data integration. We remark whether the methods are designed for paired or unpaired datasets and compare the basic network architectures and demonstrated modalities on which the respective methods have been demonstrated. Additionally, we comment on the integration tasks tackled by each model and provide a reference to the implementation.

We exclusively included methods that learn a joint embedding based on DGMs and have been demonstrated on multi-omics data of different modalities (not just, e.g., single-cell RNA-seq from different protocols).

Name	Reference	Unpaired	Network architecture	Demonstrated modalities	Integration type	Code
MAGAN	Amodio and Krishnaswamy [2]	yes	Two GANs, both unsupervised and semi-supervised training	Flow cytometry + scRNA-seq; Multiple CyTOF Panels; Multiple CyTOF Replicates	Integration of single-omics data	https://github.com/KrishnaswamyLab/MAGAN
SCIM	Stark et al. [47]	yes	multimodal autoencoders with an adversarial objective	scRNA + CyTOF, more modalities possible	Integration of multi-omics data	https://github.com/ratschlab/scim
BABEL	Wu et al. [57]	no	VAE with separate encoders and decoders, trained by cross-prediction	SNARE-seq, SHAREseq, CITE-seq, scRNA-seq, scATAC-seq	Cross-modality translation	https://github.com/wukevin/babel
Cobolt	Gong et al. [19]	yes	MVAE (direct fusion/concatenation)	SNARE-seq, 10x multiome (treated as different modalities)	Integration of multi-omics data and multi-with single-omics data	https://github.com/epurdom/cobolt
DAVAE	Hu et al. [22]	yes	VAE, shared encoder + adversarial classifier	scRNA-seq from different samples / protocols (SmartSeq2, 10X), scRNA + scATAC-seq, 10X/Visium. Requires common input features	Integration of multiple scRNA-seq into an atlas reference, transfer learning	https://github.com/jhu99/scbean
DCCA	Zuo et al. [68]	no	VAE with separate mutually supervised encoders and decoder	scRNA-seq + scATAC-seq (10x, SNARE-seq, SHARE-seq, scNMT-seq)	Transfer learning, impute missing modalities	https://github.com/cmzuo11/DCCA
MultiVI	Ashuach et al. [3]	no [1]	VAE (distributional average and penalization to mix the latent representations)	scRNA-seq + scATAC-seq (PBMC 10x)	Integration of multi-omics data and multi-omics with single-omics data, imputation of missing modalities	https://github.com/YosefLab/scvi-tools
p/mp SMILE	Xu et al. [61]	no	Modality-specific encoders trained by noise-contrastive estimation	scRNA-seq + scATAC-seq, scMethyl + scHi-C, SNARE-seq, sci-CAR, SHARE-seq, (integration of >2 modalities possible)	Integration of single-omics and multi-omics data	https://github.com/rpmccordlab/SMILE
SCALEX	Xiong et al. [58]	(yes) [2]	VAE with batch-free encoder and a batch-specific decoder	CITE-seq, spatial transcriptome MERFISH data, scRNA-seq + scATAC-seq	Integration of single-omics data, integration of multi-omics data	https://github.com/jsxlei/SCALEX
scMM	Minoura et al. [40]	no	VAE (mixture of experts)	CITE-seq + SHARE-seq	Integration of multi-omics data, cross-modal prediction	https://github.com/kodaim1115/scMM
scMVAE	Zuo and Chen [67]	no	MVAE (3 strategies: product of experts, neural network, direct concatenation)	SNARE-seq	Integration of multi-omics data	https://github.com/cmzuo11/scMVAE
TotalVI	Gayoso et al. [18]	no	VAE	CITE-seq	Integration of multi-omics data, missing protein imputation	https://github.com/YosefLab/scvi-tools
Con-AAE	Wang et al. [56]	no	Two autoencoders, using adversarial loss and latent cycle-consistency loss	sci-CAR, SNAREseq	Integration of single-omics data, integration of multi-omics data	https://github.com/kakarotcq/RNA-Seq-and-ATAC-Seq-mapping
MIRA	Lynch et al. [38]	no	VAE	SHARE-seq and 10X	Integration of multi-omics data	https://github.com/cistrome/MIRA
Polarbear	Zhang et al. [63]	yes	VAE with semi-supervised cross-domain translation	SNARE-seq (+snATAC-seq, scATAC-seq, scRNA-seq)	Cross-modality translation, align single-modality data, predict missing modalities	https://github.com/Noble-Lab/Polarbear
Multigrate	Lotfollahi et al. [34]	yes	VAE (product of experts)	CITE-seq and scRNA-seq + scATAC-seq (adaptable to other modalities)	Mapping of novel multi-omic query datasets to a reference atlas, imputation of missing modalities, integration of multi-omics	https://github.com/theislab/multigrate
Portal	Zhao et al. [65]	yes	AE+GAN: adversarial discriminators on latent spaces	Various single-cell RNA-seq (Drop-seq, 10X, SmartSeq2), scRNA (10X, DropSeq) + snRNA-seq (Split-Seq), scRNA + scATAC-seq	Integration of multi-omics and single-omics data, cross-modality translation	https://github.com/YangLabHKUST/Portal
scMVP	Li et al. [31]	no	Multimodal VAE with Gausian mixture prior and attention modules	SNARE-seq, sci-CAR, Paired-seq, SHARE-seq, 10X (could be extended to parallel profiling of other epigenomic data)	Integration of multi-omics data	https://github.com/bm2-lab/scMVP

Table 1: Overview of recently published deep learning-based methods to integrate single-cell multi-omics data.
[1]Only for mapping single-omics to multi-omics; [2]Only when converting peaks to activity scores

3.1 Approaches for paired data

The Cobolt model [19] learns shared representations between modalities and is based on a multimodal VAE, where an independent encoder network is used for each modality and the learned parameters of the posterior distributions are combined using a PoE approach. Additionally, Cobolt can jointly integrate single-modality datasets with multi-omics datasets, allowing one to draw on the many publicly available scRNA-seq or scATAC-seq datasets.

Multigrate [34] is another model that employs a PoE to combine the posteriors of different modalities. Additional datasets can be integrated into the model by minimizing the maximum mean discrepancy (MMD) loss between joint representations of different datasets.

Similar to Cobolt and Multigrate, scMM [40] is a VAE-based method that trains an encoder network for each modality independently. However, instead of combining the parameters of the posterior distributions using a PoE, a MoE is used. By equally mixing information from both modalities through the MoE, the model avoids putting too much emphasis on one individual modality only [40]. In addition, scMM provides a method for model interpretability that uses latent traversals, where synthetic cells are generated by the learned decoder and one latent variable is modified continually, while the others remain fixed. The Spearman correlations calculated between each latent variable and the features of each modality then allow relevant features to be identified. Additionally, by using a Laplace prior, scMM learns disentangled representations, with correlations between latent variables being penalized, which allows for better interpretation of individual features [50].

Similarly, the MultiVI model presented by Ashuach et al. [3] is also based on a MoE with $\alpha_m = 1/M$ where M denotes the number of modalities, as the authors use individual encoders for each data modality and then average the resulting variational posteriors. However, a regularization term is added to the ELBO, which penalizes the distance between the learned latent representations such that a joint representation can be inferred [3].

While the single-cell multi-view profiler (scMVP) [31] is also based on a multimodal VAE architecture with modality-specific encoders and decoders and a joint latent space, it more explicitly accounts for the much higher sparsity of single-cell measurements from joint profiling protocols, with a throughput of only one-tenth to one-fifth of that of single-modality assays [31]. Specifically, the authors employ attention-based building blocks for both the encoder and decoder [4, 25]. Since transformer models [52] have shown a high performance on sparse datasets in the area of natural language processing or protein structure prediction, the authors use multi-head self-attention transformer modules [52] to capture local, long-distance correlation in the encoder and decoder of the term frequency-inverse document frequency-transformed [48] scATAC-seq data. At the same time, mask attention is used in the RNA encoder and decoder. Given the latent embedding, the modality-specific decoders are weighted according to the posterior probabilities of cell-type or cluster identity. To encourage consistency of the shared latent space, the decoder-reconstructed values of each modality are again embedded into the latent space, and the KL-divergence between the joint latent embedding and the modality-specific re-embedding from the reconstructed data is minimized as an additional loss term. This corresponds to the idea of cyclical adversarial training as described in Section2.4 and Figure 1 E. More generally, this concept is based on a cycle GAN [66] and is also present in, e.g., Khan et al. [24], Wang et al. [56], Xu et al. [60], Zhao et al. [65] and Zuo et al. [68].

SCALEX [58] builds on SCALE (Single-Cell ATAC-seq Analysis via Latent feature Extraction) [59], a tool for analyzing scATAC-seq data. The developers of SCALE found that its encoder could be beneficial in disentangling cell-type- and batch-related features, which would allow for online integration of different batches. Specifically, using a VAE, SCALEX integrates different batches into a batch-invariant embedding through simultaneous learning of a batch-free encoder and a batch-specific decoder. The latter contains a domain-specific batch normalization layer. This allows the encoder to concentrate only on batch-invariant biological data components while being oblivious to batch-specific variations. The resulting generalizability of the encoder further allows for the integration of new single-cell data in an online manner, i.e., without the need to retrain the model. The authors demonstrate this property of SCALEX by generating multiple expandable single-cell atlases.

Another subgroup of models addresses the task of translating between different modalities. These cross-modality translation approaches, however, often do not learn a common latent representation of the data. For example, Polarbear [63] trains VAEs on each of two modalities (here: scRNA-seq and scATAC-seq data) and then links the respective encoders to the decoders of the other modality. The authors intend that the training in the first stage, i.e., the training of the individual VAEs, takes place on publicly available single-assay data, whereby the translation task is carried out on SNARE-seq data in a supervised manner.

Another such model called BABEL [57] similarly employs distinct modality-specific encoders and decoders for scRNA- and scATAC-seq data but utilizes a shared latent space. In contrast to PoE/MoE approaches, this joint representation is not constructed from separate spaces from each modality, but the encoders directly project onto the common latent space. Mutual cross-modal translation together with single-modality reconstruction are then used to train the model, i.e., from each modality-specific encoder, a sample of the joint latent representation is obtained and subsequently passed through both decoders to reconstruct both the scRNA and the scATAC profiles of the respective cell. Thus, both the reconstruction of the modality itself and the respective other modality based on the joint latent embedding are evaluated for each modality.

A similar approach is taken by Portal [65], where a domain translation framework is combined with an adversarial training mechanism to integrate scRNA- and scATAC-seq data. Specifically, as in [57], modality-specific encoders directly embed the data in a shared latent space and cross-modal generators are introduced to decode the latent representation to the respective other modality. The resulting domain translation networks for each modality are then trained to compete against adversarial discriminators on the domain of each modality that aims to distinguish between original cells from the respective modality and cells translated from the other modality. The discriminators are specifically designed to adaptively distinguish between domain-shared and domain-unique cells by thresholding the discriminator scores. Since, according to the authors, domain-unique cell populations are prone to be assigned with extreme discriminator scores, discriminators are, thus, made effectively inactive on cells with a high probability of being modality-specific, which avoids the risk of over-correction by enforced alignment of domain-unique cells. Further, additional regularizers are employed: an autoencoder loss based on the within-modality reconstructions, a latent alignment loss to encourage the consistency of a specific cell's embedding and the embedding of its cross-modal reconstruction, and a cosine similarity loss between cells and their cross-modal reconstructions. Notably, Portal uses the first 30 principal components of a joint PCA as inputs for the model and employs a 20-dimensional latent space, such

that the dimension reduction component is less pronounced than for the other models, and the data are not modeled as counts.

Zuo and Chen [67] have extended scMVAE and proposed Deep Cross-Omics Cycle Attention (DCCA) [68], which improves some of the weaknesses of scMVAE. DCCA combines VAEs with attention transfer. While scMVAE combines two modalities into a shared embedding, which potentially attenuates modality-specific patterns, in the case of DCCA, each data modality is processed by a separate VAE. These VAEs can then learn from each other through mutual supervision based on semantic similarity between the embeddings of each omics modality.

In the sciCAN model presented by Xu et al. [60], modality-specific autoencoders map the input data to a latent space for each modality, and a discriminator is employed to distinguish between the two modalities based on their latent representations. Additionally, a cross-modal generator is employed that generates synthetic scATAC-seq data based on the scRNA-seq latent representation, and a second discriminator is employed to distinguish between generated and real scATAC-seq samples. Additionally, the generated scATAC-seq data can be fed to the encoder again, and the latent representation is compared with the original latent representation from the scRNA-seq data used for generating the scATAC-seq data, thus introducing a cycle consistency loss (see Figure 1 E, Section 2.4). Notably, the model does not necessarily expect paired measurements from the same cell but employs a shared encoder for both modalities, and, thus, requires a common feature set.

In Zhao et al. [65], both modalities are mapped to a common latent space and cross-modal generators are employed for both modalities, together with respective discriminators for the generated samples of both modalities.

The authors of Hu et al. [22] employ an adversarial training strategy to remove batch effects, by incorporating a domain classifier that tries to determine the batch label based on the latent representation of VAE and training the VAE encoder to 'fool' the classifier via an additional adversarial loss term.

Finally, MIRA [38] combines probabilistic cell-level topic modeling [6] with gene-level regulatory potential (RP) modeling [43, 55] to determine key regulators responsible for fate decisions at lineage branch points. The topic model uses a VAE with a Dirichlet prior to learn both the topic of the gene transcription and the topic of gene accessibility for each cell to derive the cell's identity. Complementing MIRA's topic model, its RP model integrates the transcription and accessibility information for each gene locus to infer how the expression of the respective gene is influenced by surrounding regulators. To this end, the topic model learns the rate with which the regulatory influence of enhancers decays with increasing genomic distance. In addition, the identity of key regulators is identified by analyzing transcription factor motif enrichment or occupancy.

3.2 Approaches for unpaired data

Since the generation of multi-omics measurements in the same cell is still costly and experimentally complex, many methods for integrating datasets measured in different cells are being developed.

Because of the difficulty of linking latent representations learned from variational autoencoders in the absence of measurement pairing information, Lin et al. [32] proposed a transfer learning approach. Although not a DGM, it is worth mentioning in this article because of its usefulness and the possibility of adapting it to unsupervised settings. Notably, it represents a method for a horizontal alignment task,

i.e., it relies on a common set of features as anchors and thus requires the translation of scATAC peaks to gene activity scores.

In a similar spirit, the scDART model proposed by Zhang et al. [64] learns a neural network-based joint embedding or unpaired scRNA-seq and scATAC-seq data by composing the embedding network with a gene-activity module network that maps scATAC peaks to genes. In addition, scDART can leverage partial cell matching information by using it as a prior to inform the training of the gene activity function.

Similar to the sciCAN model presented by Xu et al. [60], scAEGAN [24] also embraces the concept of cycle consistency, integrating the adversarial training mechanism of a cycle GAN [66] into an autoencoder framework. Specifically, for each modality, a discriminator and a generator are defined. In addition to the standard GAN loss for each modality, a cycle loss is calculated by mapping a cell from one modality to the second modality with the second modality's generator and mapping it back to the first modality with the first modality's generator and comparing that to the original observation. Unlike for Xu et al. [60], the model does not rely on a common feature set but first trains an autoencoder model independently for each modality before training a cycle GAN on the two latent spaces to enforce their consistency.

A similar approach is employed in the Contrastive Cycle Autoencoder (Con-AAE) proposed by Wang et al. [56]. Again, the consistency between latent spaces of modality-specific autoencoders is enforced by a cycle consistency loss. However, here, it is more tightly integrated within the AE architecture, as the modality-specific encoder and decoders are used as generators, i.e., samples from one modality are embedded with the modality-specific encoder but decoded with the decoder of the other modality, and subsequently encoded with the other modality encoder back to the latent space, where they are compared with the original latent representation from the original encoder of the modality.

A purely GAN-based approach to integrating unpaired data by aligning the respective manifolds is presented in Amodio and Krishnaswamy [2].

Another line of research for the integration of unpaired multi-omics data focuses on the concept of optimal transport [41]. A separate embedding or distance matrix is constructed from each modality, and the alignment task is formulated to find an optimal coupling between the two embeddings or distance matrices. An optimal coupling corresponds to finding a map along which one modality can be "transported" with minimal cost to the other, which can be formalized as an optimal transport problem [41]. Examples for such optimal transport-based methods are UnionCom [8], SCOT [13] and Pamona [9]. While these approaches typically rely on computing a coupling between modality-specific distance matrices and are not deep learning-based, a recent approach called uniPort employs a VAE architecture and solves an optimal transport problem in the latent space. More specifically, a shared encoder that requires a common input feature set across modalities is used to project the data into a common latent space, is combined with modality-specific decoders for reconstruction, and an optimal transport loss is minimized between the latent cell embeddings from different modalities.

Finally, the recently published Graph-Linked Unified Embedding (GLUE) framework [10] is based on the construction of a guidance graph based on prior knowledge of the relations between features of the different modalities to explicitly model regulatory interactions across different modalities with distinct feature spaces. This is achieved by learning joint feature embeddings from the knowledge graph with a graph VAE and linking them to modality-specific autoencoders. Specifically, the decoder

of these modality-specific AEs is given by the inner product of the feature embeddings and the cell embeddings from the latent space of the respective modality. Additionally, the cell embeddings of different modalities are aligned using an adversarial discriminator.

4 Benchmark Dataset

To acquire an objective performance estimate of the ability of different multi-omics integration approaches to describe the biological state of a cell through learning a joint embedding from multiple modalities, we used the benchmark dataset which was provided in the course of the NeurIPS 2021 competition and for which the ground-truth cell identity labels are known [35]. This dataset was the first available multi-omics benchmarking dataset for single-cell biology. It mimics realistic challenges researchers are faced with when integrating single-cell multi-omics data, e.g., by incorporating nested donor and site batch effects [30].

Specifically, the NeurIPS benchmark dataset is a multi-donor (10 donors), multi-site (4 sites), multi-omics bone marrow dataset comprising two data types [30]:

- CITE-seq data with 81,241 cells, where for each cell RNA gene expression (GEX) and cell surface protein markers using antibody-derived tags (ADT) are jointly captured.
- 10X Multiome assay data with 62,501 cells, where nucleus GEX and chromatin accessibility measured by assay for transposase-accessible chromatin (ATAC) are jointly captured.

In total, this dataset contained information on the accessibility of 119,254 genomic regions, the expression of 15,189 genes, and the abundance of 134 surface proteins, and has been preprocessed as described in Luecken et al. [35]. We acquired the benchmark dataset from the NeurIPS 2021 website (`https://openproblems.bio/neurips_2021`), it can, however, also be accessed via `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE194122`.

As recommended in Luecken and Theis [37], we filtered this dataset for highly variable genes, as they are considered to be most informative of the variability in the data. In addition, analogous to the FindTopFeatures function of Signac [48], we filtered the ATAC data such that we retained only peaks with the 25 % highest overall counts. Finally, to determine the effect of the number of cells, we randomly subsampled the original NeurIPS dataset to subsamples containing information on 500, 1000, 2500, 5000, and 10 000 cells, where for each number of cells we sampled 10 subsamples of that size.

5 Performance metrics

Generating a highly resolved, interpretable, low-dimensional embedding capturing the underlying biological cell states is pivotal for the analysis of multi-omics data [29, 30]. We assess the performance of the compared integration approaches based on six metrics capturing the conservation of biological variation (normalized mutual information (NMI), cell type average silhouette width (ASW), trajectory conservation) and the degree of batch removal (batch ASW, site ASW, graph connectivity) [30]. These metrics are described in detail in Luecken et al. [36] and are briefly introduced below:

- **NMI** compares the overlap of two clusterings. It is used to compare the Louvain clustering of the joint embedding to the cell type labels. It ranges from 0 (uncorrelated clustering) to 1 (perfect match).
- **Cell type ASW** is used to evaluate the compactness of cell types in the joint embedding. It is based on the silhouette width, which measures the compactness of observations with the same labels. Here, the ASW was computed on cell identity labels and scaled to a value between 0 (strong misclassification) and 1 (dense and well-separated clusters).
- The **trajectory conservation** assesses the conservation of a continuous biological signal in the joint embedding. Trajectories computed using diffusion pseudotime after integration for relevant cell types are compared. Based on a diffusion map space embedding of the data, an ordering of cells in this space can be derived. Using Spearman's rank correlation coefficient between the pseudotime values before and after integration, the conservation of the trajectory can be quantified, with the scaled score ranging from 0 (reverse order of cells on the trajectory before and after integration) to 1 (same order).
- **Batch ASW** describes the ASW of batch labels per cell. The scaled score ranges from 0 to 1, where 1 indicates well-mixed batches and any deviation from 1 indicates a batch effect.
- **Site ASW** describes the ASW of site labels per cell and can be interpreted analogously to batch ASW.
- The **graph connectivity** score evaluates whether cells of the same type from different batches are close to each other in the embedding by assessing if they are all connected in this embedding's k-nearest neighbor (kNN) graph. It ranges from 0 (no cell is connected) and 1 (all cells with the same cell identity are connected).

6 Results

We use various metrics to quantify the preservation of biological variation and metrics for the removal of technical effects based on the 10-dimensional embeddings obtained when applying Cobolt, scMM, TotalVI, and SCALEX to subsamples of the NeurIPS CITE-seq dataset, and Cobolt, scMM, MultiVI, and scMVP to subsamples of the NeurIPS Multiome dataset. We randomly sampled 500, 1000, 2500, 5000, and 10000 cells ten times each and applied the models to the respective datasets. Where possible, we used the default hyperparameters of the respective models (Supplementary: Hyperparameters). When applying scMM to the CITE-seq data, we frequently observed non-converging training runs in particular for larger sample sizes and subsequently lowered the learning rate for sample sizes > 2500. However, training still did not converge in one dataset of size 2500 and 5000 each and 4 datasets of size 10,000, which were omitted from further analysis. In general, similar performances were achieved irrespective of which of the two data types we used for deriving a joint embedding (**Figures 4 and 5**).

We compare our results with the metric values achieved by the models of the NeurIPS 2021 competition for the integration of the Multiome dataset (data points were extracted via WebPlotDigitizer-4.5 [44] from Supplementary Figure 6 of [30]). However, as we merely used a subset of at most 10,000

cells of the original benchmark dataset, we expect our investigated algorithms to score higher for most metrics if they were to be subjected to the complete benchmark dataset.

By visual inspection of the Uniform Manifold Approximation and Projection (UMAP) (Becht et al. [5], Konopka and Konopka [28] version 0.2.8.0 with default parameters) plots of one exemplary subsample (**Figure 2** and **Figure 3**), we see that MultiVI shows no obvious clustering for 500 cells (2, top panel). In contrast, defined cluster structures are beginning to build at this low cell number, and become more refined for 10,000 cells, for all other investigated tools. This behavior of MultiVI for smaller numbers of cells is also reflected in lower values for most of the investigated performance metrics (**Figures 2 and 3**). Interestingly, the TotalVI tool, which is built on a similar architecture and was used for the CITE-seq dataset does not show such behavior (4, top panel).

UMAP plots including further meta information on the embedded cells are given in **Supplementary Figures 3 to 18** for the exemplary subsample.

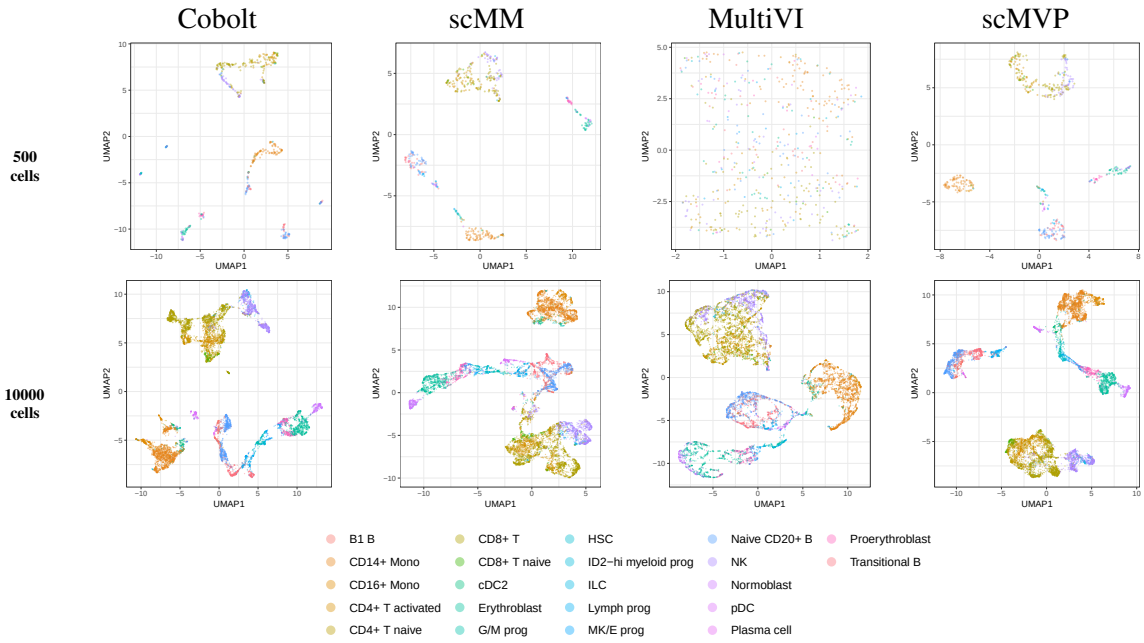

Figure 2: UMAP of the 10-dimensional latent space of Cobolt, scMM, MultiVI, and scMVP based on 500 (top) and 10000 (bottom) cells of one exemplary subsample from the Multiome dataset each. The color coding corresponds to manually annotated cell types as provided by Luecken et al. [35].

6.1 Preserving biological information

We assess the preservation of biological variation based on the NMI, cell type ASW, and the trajectory conservation scores (**Figure 4**).

NMI, as a measure of cluster overlap, reaches values of approx. 0.7 for all Multiome and CITE-seq integrating models. The NMI is slightly lower than what was achieved during the NeurIPS 2021 competition, where the best competition entries reached an NMI of close to 0.8 for the complete Multiome dataset [30].This is to be expected as we evaluate the models in a low sample size scenario. MultiVI profits greatly from a larger cell number, while an increasing cell number only slightly increases the performance of the other models. Across most sample sizes, Cobolt performed best, both for the

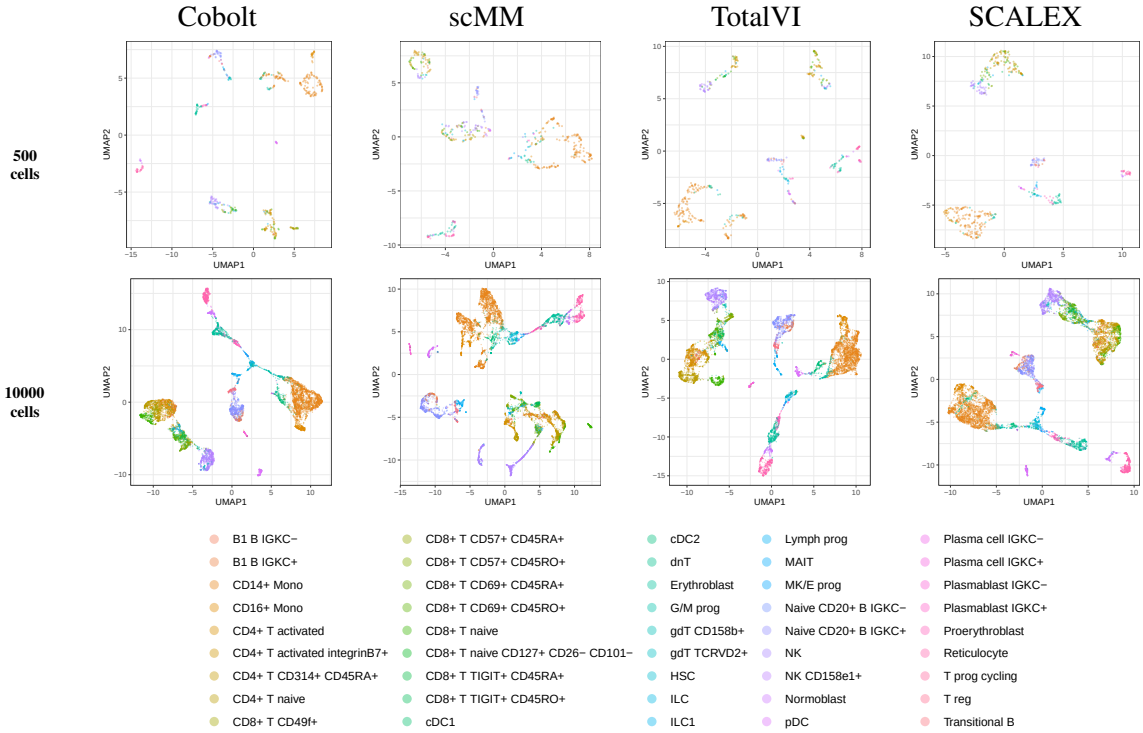

Figure 3: UMAP of the 10-dimensional latent space of Cobolt, scMM, TotalVI, and SCALEX based on 500 (top) and 10000 (bottom) cells of one exemplary subsample from the CITE-seq dataset each. The color coding corresponds to manually annotated cell types as provided by Luecken et al. [35]. The following cell types are not present in the 500 cell sample: CD4+ T CD314+ CD45RA+, CD8+ T naive CD127+ CD26- CD101-, cDC1, dnT, Plasma cell IGKC-, Plasma cell IGKC+, Plasmablast IGKC-, T prog cycling.

Multiome and CITE-seq datasets. Only for larger cell numbers scMVP performs better than Cobolt.

Cell type ASW is a measure of cluster compactness and overlap. We see values of around 0.5 for the Multiome and CITE-seq datasets, which implies overlapping of clusters and only a moderate separation. This is slightly lower than the 0.6 that models have achieved in the NeurIPS 2021 competition [30]. For Multiome data, the impact of cell numbers was minor in Cobolt, scMM, and scMVP representations and higher for MultiVI. For CITE-seq data, only scMM and TotalVI show a dependence between cell type ASW and cell number. As expected, increasing the number of cells leads to a decrease in variance.

The trajectory conversation score measures the preservation of a biological signal, e.g. in the form of developmental processes. For the CITE-seq dataset, all models reach comparable scores of around 0.9 irrespective of the cell numbers, with a substantial decrease in variance for larger cell numbers. In contrast, for the Multiome dataset, an increase in cell numbers affects the trajectory conservation score. While this impact is minor for Cobolt, scMM, and scMVP, with Cobolt performing best throughout, MultiVI shows a surge with increasing cell numbers. The maximum score that models reach in our analysis slightly exceeds the median of the trajectory conservation scores of around 0.9 achieved by models of the NeurIPS 2021 competition [30].

Taken together, Cobolt is the strongest performing model based on almost all biology preservation metrics and across the two data types, performing well even in scenarios with small sample sizes (**Supplementary Figure 1**).

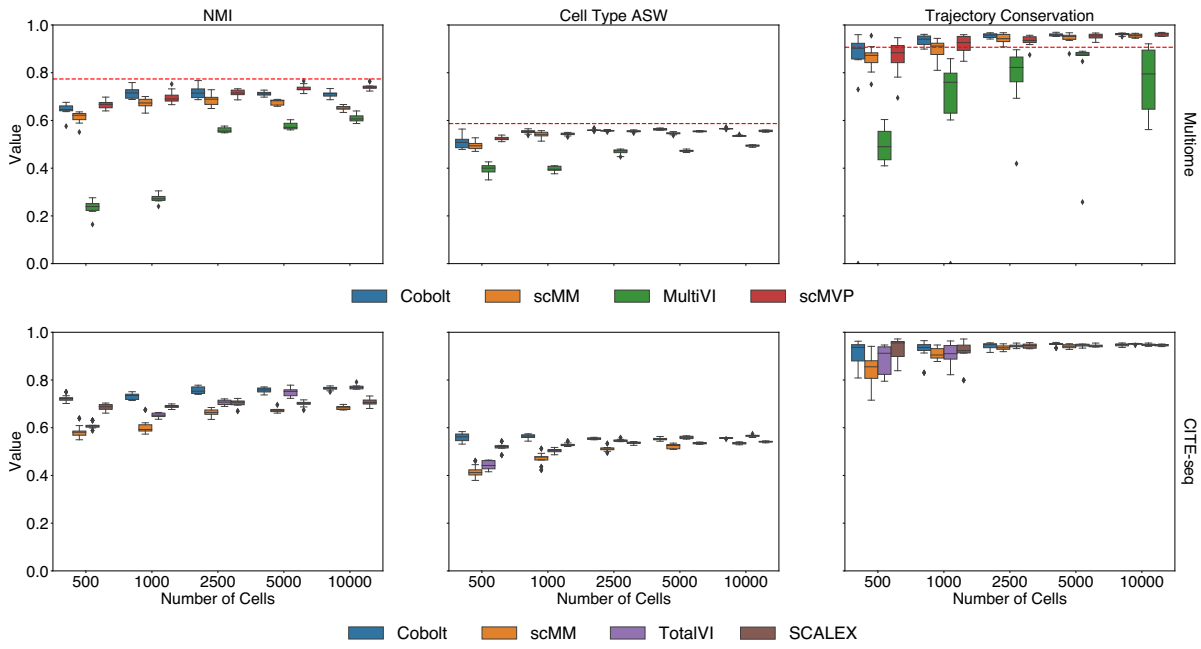

Figure 4: **Biological preservation metrics.** NMI, cell type ASW, and trajectory conservation score indicate the biological preservation quality achieved by joint embeddings from various models. On the 10x Multiome data, the performance of Cobolt, scMM, MultiVI, and scMVP is shown (top), whereas the performance on CITE-seq data is shown for Cobolt, scMM, TotalVI, and SCALEX (bottom). The red dashed lines represent the median performance across all models (excluding the ones with a metric equal to zero) presented and available in the NeurIPS 2021 Multimodal data integration competition [30].

6.2 Removing technical effects

We assess the removal of technical artifacts based on the batch ASW and graph connectivity score (**Figure 5**). As a measure of between-site technical variation and to account for the shortcomings of batch ASW (which does not sufficiently account for the nested batch effects of donors and sites) and graph connectivity (which is not sufficiently challenging) [30], we also assess batch ASW with the site as a covariate (‘Site ASW‘), as has been suggested by Lance et al. [30].

The Batch ASW score of around 0.8 that we observe in our results indicates only a minor batch effect, although the score is slightly lower than the 0.9 that models achieved in the course of the NeurIPS 2021 competition [30]. There is a slight increase in performance for increasing cell numbers across both datasets. For the Multiome dataset, scMVP consistently performed best, while MultiVI scored lowest for most cell number settings. For the CITE-seq dataset, SCALEX shows the highest Batch ASW score across all cell number settings, implying superior handling of batch effects even with small sample sizes. This is in line with SCALEX being specifically designed to separate batch-related from batch-invariant components [58].

The graph connectivity score indicates how well cells of the same cell type and cells coming from different batches are connected in the joint embedding. For the Multiome dataset, MultiVI’s graph connectivity score is considerably lower for small sample sizes, while all models improve performance with an increasing number of cells. With the best performing models reaching a score of almost 1 for

higher cell numbers in the case of the Multiome dataset, this is in line with the scores achieved by the models of the NeurIPS competition [30]. For the CITE-seq dataset, the performance of TotalVI increased with increasing cell numbers, achieving the highest graph connectivity score for 5,000 and 10,000 cells. In contrast, the number of cells had only a minor effect on the other models. scMM consistently had the lowest graph connectivity score for the CITE-seq dataset.

Site ASW captures site-specific batch effects. Compared to Batch ASW, the performance differences between the models that we applied to the CITE-seq dataset are enhanced. For the CITE-seq dataset, Cobolt and SCALEX perform best, with Cobolt surpassing SCALEX for increasing cell numbers. For the Multiome data, the spread of the investigated models is comparable to the one of Batch ASW. scMVP achieved the highest Site ASW scores, which is in agreement with its high Batch ASW score.

scMVP and SCALEX are among the best performing models for metrics considering the removal of technical effects, whereas MultiVI's performance suffers, especially in scenarios with small numbers of cells (**Supplementary Figure 2**).

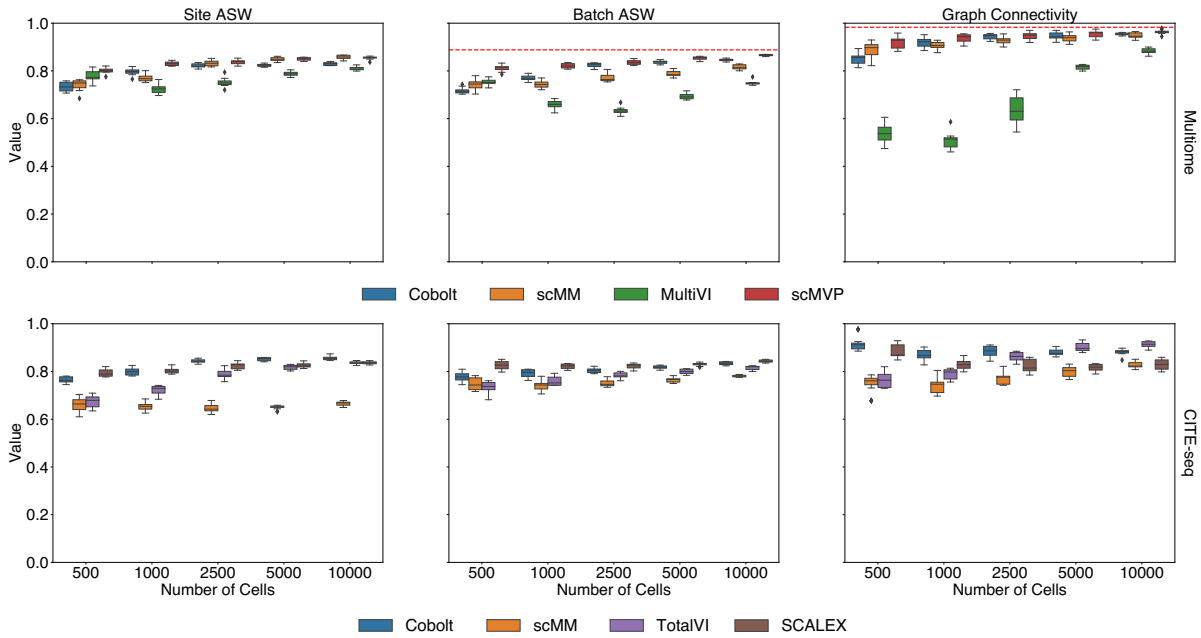

Figure 5: **Technical effect removal metrics**. Site ASW, batch ASW, and graph connectivity score indicate the quality of technical effects removal achieved by joint embeddings from various models. On the 10x Multiome data, the performance of Cobolt, scMM, MultiVI, and scMVP is shown (top), whereas the performance on CITE-seq data is shown for the Cobolt, scMM, TotalVI, and SCALEX (bottom). The red dashed lines represent the median performance across all models (excluding the ones with a metric equal to zero) presented and available in the NeurIPS 2021 Multimodal data integration competition [30].

6.3 Usability

We experienced memory problems when running MultiVI, which was also observed by Zhang et al. [63]. However, as this is a preprint and the model is embedded in scvi-tools[16], we suspect that this problem might be fixed in the near future.

In addition, when trying to apply the domain-adversarial VAE from Hu et al. [22], we were unable to correctly install the implementation of the approach provided by the authors via the Python package `scbean`.

The scMM model by [40] was easily usable. The authors provide both a command-line interface and a script that is straightforward to adapt and run. However, HDF5-based data (such as the popular 'anndata' objects) has to be manually restructured to separate files to be used as input for the model. For CITE-seq-data, model training frequently did not converge, in particular for larger sample sizes, which could only partly be improved by lowering the learning rate. While this behavior did not occur with a very small learning rate (2 orders of magnitude smaller than the default used by Minoura et al. [40]), this also substantially lowered the performance.

To run the scMVP model by [31], package dependency issues had to be resolved manually. Here, too, data had to be restructured manually to fit the custom input data structs defined by the authors. Adapting and running the model and extracting the learned embedding was straightforward.

All in all, the tools we used were relatively easy to use and adapt, though in most cases not without at least intermediate programming skills (e.g., to transform own data into very specific and largely undocumented data structs defined by the authors).

7 Outlook/Discussion

The rapid emergence of experimental protocols for profiling several omics layers from the same cell or in independent experiments is closely followed by the development of corresponding computational models for analyzing and integrating such data. These methods promise to answer biological questions previously out of reach. Still, they have so far been hampered by often rather small and sparse datasets and the lack of a systematic overview and comparison. In particular, considering the sparsity and high dimensionality inherent to single-cell (multi-)omics data, researchers seek to identify a low-dimensional embedding that integrates the information from multiple modalities and can be used for further downstream analysis. Consequently, many computational tools to infer such a joint latent representation have recently been proposed, often based on deep learning approaches due to their success in identifying complex structures from data in unsupervised settings. Specifically, deep generative models such as VAEs that infer a low-dimensional, compressed representation of the input data in an unsupervised way are among the most popular solutions, often including additional components or custom architectures to accommodate the properties of single-cell multi-omics data and facilitate specific characteristics of the learned embedding.

Due to the rapidly growing number of complex methodological proposals for solving the challenging task of computationally integrating multi-omics data, an overview and categorization of such models are essential in understanding the advantages and disadvantages of the different methods. We have compiled a comprehensive review of the literature on DGMs for learning joint embeddings of multi-omics data and categorized the different models according to their architectural choices.

In addition to this overview, we have also illustrated the robustness of selected models to small sample sizes, where sample size refers to the number of cells in the dataset. For evaluating model performance, we have relied on the guidelines of a comprehensive benchmarking project [35]. We have evaluated the models with established metrics concerning their ability to adjust for technical

effects while maintaining biological signals. Our analyses have shown that Cobolt, an approach that uses a multimodal VAE with products of experts to combine individual embeddings, delivers the best performance for most biological preservation metrics, particularly for small numbers of cells. On the other hand, scMVP, an approach that employs attention-based components and a dedicated architecture to deal with the sparsity of scATAC-seq data, and SCALEX score highest for metrics related to removing technical artifacts.

In this work, we, however, merely discussed some of all available omics modalities, and the performance of the models may be affected for the better or the worse if applied to other data types due to differing data characteristics, e.g., in the degree of sparsity.

Interpretability is an aspect that is of great importance for the application of DGMs [50]. Some of the models we have reviewed already offer the possibility of making the corresponding outputs interpretable for users. For example, post-hoc methods such as applying archetypal analysis [12] to the joint embedding as conducted by TotalVI [18], can make the models explainable after they have been trained. On the other hand, model-based interpretability can be directly incorporated into the model architecture to allow for immediate interpretation, such as the latent traversals and specification of a dedicated prior to facilitate disentanglement in [40]. However, no dominant approach has yet emerged in this area, providing scope for new developments.

Especially for users with little programming experience, some of the models investigated will be difficult to apply, as they require, e.g., the use of command-line tools. Here, libraries such as scvi-tools [17] offer a significant benefit by providing extensive documentation and exemplary applications.

In the future, linking information from measurements of transcriptomes, epigenomes, proteomes, chromatin organization, etc., could lead to a deeper understanding of cellular processes. Scientists could then further enhance their understanding of these processes by information on the spatial context.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

E.B., M.H., and M.T. conceived the idea for the manuscript, conducted the analyses, and wrote the manuscript. C.K. and H.B. contributed to writing and proofread the manuscript. All authors read and approved the final manuscript.

Funding

The work of M.H. is funded by the DFG (German Research Foundation) – 322977937/GRK2344. M.T. and H.B. are supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) — Project-ID 431984000 - SFB 1453. C.K. is funded by the German Ministry of Education and Research by grant EA:Sys[FKZ031L0080]. C.K. and E.B. are funded by the Deutsche

Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (CIBSS-EXC-2189-2100249960-390939984).

Supplemental Data

Data Availability Statement

The benchmark dataset can be accessed via `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE194122`.

The joint embeddings learned by all models can be found at `https://doi.org/10.5281/zenodo.6616542`.

The code to reproduce all analyses and figures can be found at `https://github.com/MTreppner/multiomics_dgms`.

References

[1] Adossa, N., Khan, S., Rytkönen, K. T., and Elo, L. L. (2021). Computational strategies for single-cell multi-omics integration. *Computational and Structural Biotechnology Journal*

[2] Amodio, M. and Krishnaswamy, S. (2018). Magan: Aligning biological manifolds. In *International Conference on Machine Learning* (PMLR), 215–223

[3] Ashuach, T., Gabitto, M. I., Jordan, M. I., and Yosef, N. (2021). Multivi: deep generative model for the integration of multi-modal data. *bioRxiv*

[4] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*

[5] Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W., Ng, L. G., et al. (2019). Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology* 37, 38–44

[6] Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM* 55, 77–84

[7] Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association* 112, 859–877

[8] Cao, K., Bai, X., Hong, Y., and Wan, L. (2020). Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics* 36, i48–i56. doi:10.1093/bioinformatics/btaa443

[9] Cao, K., Hong, Y., and Wan, L. (2021). Manifold alignment for heterogeneous single-cell multi-omics data integration using pamona. *Bioinformatics* 38, 211–219. doi:10.1093/bioinformatics/btab594

[10] Cao, Z.-J. and Gao, G. (2022). Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*

[11] Colomé-Tatché, M. and Theis, F. J. (2018). Statistical single cell multi-omics integration. *Current Opinion in Systems Biology* 7, 54–59

[12] Cutler, A. and Breiman, L. (1994). Archetypal analysis. *Technometrics* 36, 338–347

[13] Demetci, P., Santorella, R., Sandstede, B., Noble, W. S., and Singh, R. (2022). Scot: Single-cell multi-omics alignment with optimal transport. *Journal of Computational Biology* 29, 3–18. doi:10.1089/cmb.2021.0446

[14] Ding, J., Condon, A., and Shah, S. P. (2018). Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nature communications* 9, 1–13

[15] Erfanian, N., Heydari, A. A., Iañez, P., Derakhshani, A., Ghasemigol, M., Farahpour, M., et al. (2021). Deep learning applications in single-cell omics data analysis. *bioRxiv*

[16] Gayoso, A., Lopez, R., Xing, G., Boyeau, P., Valiollah Pour Amiri, V., Hong, J., et al. (2022). A python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology* doi:10.1038/s41587-021-01206-w

[17] Gayoso, A., Lopez, R., Xing, G., Boyeau, P., Wu, K., Jayasuriya, M., et al. (2021). Scvi-tools: A library for deep probabilistic analysis of single-cell omics data. *bioRxiv*

[18] Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K. L., Streets, A., et al. (2021). Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nature Methods* 18, 272–282

[19] Gong, B., Zhou, Y., and Purdom, E. (2021). Cobolt: integrative analysis of multimodal single-cell sequencing data. *Genome biology* 22, 1–21

[20] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. *Advances in neural information processing systems* 27

[21] Grapov, D., Fahrmann, J., Wanichthanarak, K., and Khoomrung, S. (2018). Rise of deep learning for genomic, proteomic, and metabolomic data integration in precision medicine. *Omics: a journal of integrative biology* 22, 630–636

[22] Hu, J., Zhong, Y., and Shang, X. (2021). A versatile and scalable single-cell data integration algorithm based on domain-adversarial and variational approximation. *Briefings in Bioinformatics* 23. doi:10.1093/bib/bbab400

[23] Kang, M., Ko, E., and Mersha, T. B. (2021). A roadmap for multi-omics data integration using deep learning. *Briefings in Bioinformatics*

[24] Khan, S. A., Lehmann, R., Martinez-de Morentin, X., Ruiz, A. M., Lagani, V., Kiani, N. A., et al. (2022). scaegan: Unification of single-cell genomics data by adversarial learning of latent space correspondences. *bioRxiv* doi:10.1101/2022.04.19.488745

[25] Kim, Y., Denton, C., Hoang, L., and Rush, A. M. (2017). Structured attention networks. *International Conference on Learning Representations (ICLR)*

[26] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*

[27] Kingma, D. P. and Welling, M. (2019). An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*

[28] Konopka, T. and Konopka, M. T. (2018). R-package: umap. *Uniform Manifold Approximation and Projection*

[29] Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., et al. (2020). Eleven grand challenges in single-cell data science. *Genome biology* 21, 1–35

[30] Lance, C., Luecken, M. D., Burkhardt, D. B., Cannoodt, R., Rautenstrauch, P., Laddach, A. C., et al. (2022). Multimodal single cell data integration challenge: results and lessons learned. *bioRxiv*

[31] Li, G., Fu, S., Wang, S., Zhu, C., Duan, B., Tang, C., et al. (2022). A deep generative model for multi-view profiling of single-cell rna-seq and atac-seq data. *Genome biology* 23, 1–23

[32] Lin, Y., Wu, T.-Y., Wan, S., Yang, J. Y., Wong, W. H., and Wang, Y. (2022). scjoint integrates atlas-scale single-cell rna-seq and atac-seq data with transfer learning. *Nature Biotechnology* , 1–8

[33] Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature methods* 15, 1053–1058

[34] Lotfollahi, M., Litinetskaya, A., and Theis, F. J. (2022). Multigrate: single-cell multi-omic data integration. *bioRxiv*

[35] Luecken, M. D., Burkhardt, D. B., Cannoodt, R., Lance, C., Agrawal, A., Aliee, H., et al. (2021). A sandbox for prediction and integration of dna, rna, and proteins in single cells. *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*

[36] Luecken, M. D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M. F., et al. (2021). Benchmarking atlas-level data integration in single-cell genomics. *Nature methods* , 1–10

[37] Luecken, M. D. and Theis, F. J. (2019). Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology* 15, e8746

[38] Lynch, A. W., Theodoris, C. V., Long, H., Brown, M., Liu, X. S., and Meyer, C. A. (2022). Mira: Joint regulatory modeling of multimodal expression and chromatin accessibility in single cells. *bioRxiv* doi:10.1101/2021.12.06.471401

[39] Marouf, M., Machart, P., Bansal, V., Kilian, C., Magruder, D. S., Krebs, C. F., et al. (2020). Realistic in silico generation and augmentation of single-cell rna-seq data using generative adversarial networks. *Nature communications* 11, 1–12

[40] Minoura, K., Abe, K., Nam, H., Nishikawa, H., and Shimamura, T. (2021). A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. *Cell Reports Methods* 1, 100071

[41] Peyré, G. and Cuturi, M. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning* 11, 355–607. doi:10.1561/2200000073

[42] Picard, M., Scott-Boyer, M.-P., Bodein, A., Périn, O., and Droit, A. (2021). Integration strategies of multi-omics data for machine learning analysis. *Computational and Structural Biotechnology Journal*

[43] Qin, Q., Fan, J., Zheng, R., Wan, C., Mei, S., Wu, Q., et al. (2020). Lisa: inferring transcriptional regulators through integrative modeling of public chromatin accessibility and chip-seq data. *Genome biology* 21, 1–14

[44] [Dataset] Rohatgi, A. (2021). Webplotdigitizer: Version 4.5

[45] Shi, Y., Paige, B., Torr, P., et al. (2019). Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in Neural Information Processing Systems* 32

[46] Stanojevic, S., Li, Y., and Garmire, L. X. (2022). Computational methods for single-cell multi-omics integration and alignment. *arXiv preprint arXiv:2201.06725*

[47] Stark, S. G., Ficek, J., Locatello, F., Bonilla, X., Chevrier, S., Singer, F., et al. (2020). SCIM: universal single-cell matching with unpaired feature sets. *Bioinformatics* 36, i919–i927. doi: 10.1093/bioinformatics/btaa843

[48] Stuart, T., Srivastava, A., Madad, S., Lareau, C. A., and Satija, R. (2021). Single-cell chromatin state analysis with signac. *Nature methods* 18, 1333–1341

[49] Tarazona, S., Arzalluz-Luque, A., and Conesa, A. (2021). Undisclosed, unmet and neglected challenges in multi-omics studies. *Nature Computational Science* , 1–8

[50] Treppner, M., Binder, H., and Hess, M. (2022). Interpretable generative deep learning: an illustration with single cell gene expression data. *Hum Genet* , 1–18

[51] Treppner, M., Salas-Bastos, A., Hess, M., Lenz, S., Vogel, T., and Binder, H. (2021). Synthetic single cell rna sequencing data from small pilot studies using deep generative models. *Scientific reports* 11, 1–11

[52] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, eds. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc.), vol. 30

[53] Veenstra, T. D. (2021). Omics in systems biology: Current progress and future outlook. *Proteomics* 21, 2000235

[54] Veenstra, T. D. (2021). Systems biology and multi-omics. *Proteomics* 21, 2000306

[55] Wang, S., Sun, H., Ma, J., Zang, C., Wang, C., Wang, J., et al. (2013). Target analysis by integration of transcriptome and chip-seq data with beta. *Nature protocols* 8, 2502–2515

[56] Wang, X., Hu, Z., Yu, T., Wang, Y., Wang, R., Wei, Y., et al. (2022). Contrastive cycle adversarial autoencoders for single-cell multi-omics alignment and integration. *bioRxiv* doi:10.1101/2021.12.12.472268

[57] Wu, K. E., Yost, K. E., Chang, H. Y., and Zou, J. (2021). Babel enables cross-modality translation between multiomic profiles at single-cell resolution. *Proceedings of the National Academy of Sciences* 118

[58] Xiong, L., Tian, K., Li, Y., and Zhang, Q. C. (2021). Online single-cell data integration through projecting heterogeneous datasets into a common cell-embedding space. *bioRxiv*

[59] Xiong, L., Xu, K., Tian, K., Shao, Y., Tang, L., Gao, G., et al. (2019). Scale method for single-cell atac-seq analysis via latent feature extraction. *Nature communications* 10, 1–10

[60] Xu, Y., Begoli, E., and McCord, R. P. (2021). scican: Single-cell chromatin accessibility and gene expression data integration via cycle-consistent adversarial network. *bioRxiv* doi:10.1101/2021.11.30.470677

[61] Xu, Y., Das, P., and McCord, R. P. (2021). SMILE: mutual information learning for integration of single-cell omics data. *Bioinformatics* 38, 476–486. doi:10.1093/bioinformatics/btab706

[62] Xu, Y., Zhang, Z., You, L., Liu, J., Fan, Z., and Zhou, X. (2020). scigans: single-cell rna-seq imputation using generative adversarial networks. *Nucleic acids research* 48, e85–e85

[63] Zhang, R., Meng-Papaxanthos, L., Vert, J.-P., and Noble, W. S. (2022). Semi-supervised single-cell cross-modality translation using polarbear. In *Research in Computational Molecular Biology*, ed. I. Pe'er (Cham: Springer International Publishing), 20–35

[64] Zhang, Z., Yang, C., and Zhang, X. (2022). Integrating unmatched scrna-seq and scatac-seq data and learning cross-modality relationship simultaneously. *bioRxiv* doi:10.1101/2021.04.16.440230

[65] Zhao, J., Wang, G., Ming, J., Lin, Z., Wang, Y., Consortium, T. T. M., et al. (2022). Adversarial domain translation networks for fast and accurate integration of large-scale atlas-level single-cell datasets. *bioRxiv* doi:10.1101/2021.11.16.468892

[66] Zhu, J.-Y., Park, T., Isola, P., , and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *IEEE International Conference on Computer Vision (ICCV)*

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

[68] Zuo, C., Dai, H., and Chen, L. (2021). Deep cross-omics cycle attention model for joint analysis of single-cell multi-omics data. *Bioinformatics* 37, 4091–4099