TarDis: Achieving Robust and Structured Disentanglement of Multiple Covariates

Kemal Inecik [1 2] **Aleyna Kara** [1 3] **Antony Rose** [4 5] **Muzlifah Haniffa** [4 5] **Fabian J. Theis** [1 6]

Abstract

Addressing challenges in domain invariance within single-cell genomics necessitates innovative strategies to manage the heterogeneity of multi-source datasets while maintaining the integrity of biological signals. We introduce *TarDis*, a novel deep generative model designed to disentangle intricate covariate structures across diverse biological datasets, distinguishing technical artifacts from true biological variations. By employing tailored covariate-specific loss components and a self-supervised approach, *TarDis* effectively generates multiple latent space representations that capture each continuous and categorical target covariate separately, along with unexplained variation. Our extensive evaluations demonstrate that *TarDis* outperforms existing methods in data integration, covariate disentanglement, and robust out-of-distribution predictions. The model's capacity to produce interpretable and structured latent spaces, including ordered latent representations for continuous covariates, enhances its utility in hypothesis-driven research. Consequently, *TarDis* offers a promising analytical platform for advancing scientific discovery, providing insights into cellular dynamics, and enabling targeted therapeutic interventions.

1. Introduction

Domain invariance tackles the challenge of learning from datasets that, while representing the same physical phenomena, originate from disparate sources such as different users, acquisition devices, or locations (Andéol et al., 2023). As the data source often lacks direct relevance to the task, the objective is to develop a model that maintains performance robustness by being invariant to these domain variations. This invariance not only enhances model reliability across shifts, whether subpopulational (Koh et al., 2021) or distributional (Goel et al., 2020), but also is an end in itself where the source is obscured to comply with data protection requirements (Hajihassnai et al., 2021). Such shifts, frequently observed in practical machine learning scenarios, necessitate models to be resilient to variations in multi-domain datasets by learning to minimize the disparity in data distributions within the representation space; ideally achieving a low metric distance between them. This concept is closely aligned with distributionally robust optimization strategies, promoting the development of universally applicable machine learning models that withstand out-of-distribution variations (Lu et al., 2021a; Yin et al., 2021; Guo et al., 2024; Sturma et al., 2024).

The identification of spurious correlations within these multi-domain datasets can provide critical insights for certain downstream applications, enriching the interpretive scope beyond mere domain invariance. Moreover, models leveraging data representations or predictors derived from true correlations, including domain-specific attributes or nuisance factors, more effectively discern causal relationships, thereby enhancing their generalization capabilities (Ahuja et al., 2020; Aliee et al., 2023). This recognition has spurred interest in disentangled representation learning, aiming to segregate and independently model spurious and invariant characteristics within the data (Aliee et al., 2023; Arjovsky et al., 2019; Kong et al., 2022). Developing invariant representation learning models is a complex multi-objective optimization problem, frequently necessitating linear constraints on the data representations and classifiers (Ahuja et al., 2020; Kong et al., 2022; Ahuja et al., 2021), or the incorporation of conditional priors within the VAE framework (Aliee et al., 2023; Lu et al., 2021b).

Existing invariant representation learning methods often fail for continuous domain problems, an area that is significantly underexplored yet critically important (Yong et al., 2024;

[1]Institute of Computational Biology, Helmholtz Center Munich, Neuherberg, Germany [2]School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany [3]Department of Computer Science, Technical University of Munich, Garching, Germany [4]Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, UK [5]Biosciences Institute, Newcastle University, Newcastle upon Tyne, UK [6]Department of Mathematics, Technical University of Munich, Garching, Germany. Correspondence to: Fabian Theis <fabian.theis@helmholtz-munich.de>.

3rd AI4Science Workshop in the 41^{st} International Conference on Machine Learning (ICML), 2024.

Azzam et al., 2021; Zhang & Davison, 2021). Examples include patient monitoring systems where physiological spurious data varies daily and across activities (Cao et al., 2023), finance, where models predicting stock prices or market trends must generalize across varying economic conditions and times (Huang et al., 2023), and climate modeling, where models use invariant learning to forecast weather or long-term climate changes across diverse locations and time periods (Beucler et al., 2024). Existing methods are generally designed for discrete categorical domains and struggle with the continuous nature of many real-world tasks. This leads to challenges such as sparse data in each domain, making it hard to accurately estimate invariant correlations, and segmentation of continuous data into discrete blocks which can misrepresent true data distributions. Addressing these issues is crucial for advancing model robustness and ensuring applicability in dynamic environments.

In the context of domain invariance, multi-domain and multi-condition single-cell genomics datasets present a critical testbed where the integration of data representations confronts complex challenges in biological and pharmaceutical research (Heumos et al., 2023). Single-cell genomics offers a granular view of individual cells' genetic diversity, highlighting the variability among cells and essential for understanding cellular and molecular processes (Inecik & Theis, 2023; Perez et al., 2022; Bergen et al., 2020). However, the data often come from a range of labs and varied experimental setups, incorporating batch effects and technical artifacts that can mask true biological signals (Eraslan et al., 2019; Lopez et al., 2018). These challenges are compounded when data includes cells affected by chemical or genetic perturbations, sourced from diseased states, or differing in their origin, such as specific organs, organisms, developmental stages, ethnicity, age, sex and other factors that further contribute to variability (Srivatsan et al., 2020; Sikkema et al., 2023; Hrovatin et al., 2023; Haniffa et al., 2021; Muus et al., 2021). Effective data integration is vital for separating technical artifacts from relevant biological signals, facilitating a robust comparison of biological landscapes across various domains and enhancing our understanding of the underlying cellular dynamics, with significant implications for advancing disease research and therapeutic development (Regev et al., 2017; Rood et al., 2022).

Hence, it becomes essential to disentangle invariant and spurious correlations for single-cell data integration, where spurious correlations often obscure biological signals. The disentanglement of these elements not only enhances data integration by clarifying underlying biological processes but also bolsters out-of-distribution (OOD) prediction capabilities (Aliee et al., 2023; Liu et al., 2024). Furthermore, there is a compelling need for researchers to explore the potential effects of one covariate on another, whether categorical or continuous, by manipulating such disentangled latent representations. For instance, adjusting the continuous *'drug dose'* representation while holding the representations of *'disease state'*, *'patient'*, and continuous *'age'* constant could reveal the dose-dependent effects on gene expression independent of the disease's progression or patient characteristics. Such analyses would deepen our understanding of the interactions between various factors at the cellular level, thereby unlocking new avenues for complex, hypothesis-driven research with single-cell genomics data.

To address the complexities inherent in multi-domain and multi-condition datasets, we introduce *TarDis*, a novel end-to-end deep generative model specifically designed for the *tar*geted *dis*entanglement of multiple covariates, such as those encountered in extensive single-cell genomics data. [1] *TarDis* employs covariate-specific loss functions through a self-supervision strategy, enabling the learning of disentangled representations that achieve accurate reconstructions and effectively preserve essential biological variations across diverse datasets. It eschews additional architectural complexities, enabling straightforward application to large datasets. *TarDis* ensures the independence of invariant signals from noise, enhancing interpretability that is crucial for extracting biological insights obscured by spurious data correlations. *TarDis* handles both categorical and, notably, continuous variables, demonstrating its adaptability to diverse data characteristics and allowing for a granular understanding and representation of underlying data dynamics within a coherent and interpretable latent space. This capability is instrumental for exploring complex biological phenomena and conducting hypothesis-driven research. Empirical benchmarking across multiple datasets highlight *TarDis*'s superior performance in covariate disentanglement, data integration, and out-of-distribution predictions, significantly outperforming existing models. [2] [3]

2. Method

Let $\mathcal{D}$ represent a single-cell genomics dataset containing N_C cells, where each cell, denoted as n, is characterized by its gene expression ($\mathbf{x}_n$) and associated covariates ($\mathbf{s}_n$). The gene expression is represented by a count vector $\mathbf{x}_n = [x_{ng}]_{g=1}^{N_G}$, where $x_{ng} \in \mathbb{Z}_{\geq 0}$ is the expression count of gene g, and N_G is the total number of genes in the dataset. Additionally, each cell n is associated with a vector of covariates $\mathbf{s}_n = [s_{nk}]_{k=1}^{N_K}$, which may be either continuous or discrete, and N_K indicates the number of covariates. *TarDis* constructs a latent representation $\mathbf{z}_n$ for gene expression $\mathbf{x}_n$, organized as $\mathbf{z}_n = (\mathbf{z}_{n0}, [\mathbf{z}_{nk}]_{k \in J_k})$, where $J_k \subseteq \{1, \ldots, N_K\}$ denotes the subset of covariates

[1] *"A place for everything, and everything in its place."* — Benjamin Franklin

[2] See Appendix A for discussions regarding relevant works.

[3] Source code is on GitHub, under `theislab/tardis`.

targeted for disentanglement. Specifically, $\mathbf{z}_{nk}$ is a latent vector constructed for each targeted covariate, while $\mathbf{z}_{n0}$ captures residual information independent of targeted covariates. During model training, *TarDis* employs a novel approach to foster disentanglement by generating pairs of additional latent vectors $(\mathbf{z}_n^{(k)})^-$ and $(\mathbf{z}_n^{(k)})^+$ corresponding to two data points $(\mathbf{x}_n^{(k)})^-$ and $(\mathbf{x}_n^{(k)})^+$. These data points are selected *randomly* and differ in the k_{th} covariate value, such that $(\mathbf{s}_{nk}^{(k)})^+ = \mathbf{s}_{nk}$ and $(\mathbf{s}_{nk}^{(k)})^- \neq \mathbf{s}_{nk}$.

The primary objective of *TarDis* training is to optimize the latent vectors based on a distance measure F. While F is defined conceptually as a real-valued function, $F : \mathbb{R}^{|\mathbf{z}_{nk}|} \to \mathbb{R}_{\geq 0}$, here just to illustrate the underlying concept, practical implementation typically employ multiple loss terms instead of a single function for optimizing latent vectors, as will be discussed in further detail. For each covariate $k \in J_k$, F should satisfy $F(\mathbf{z}_{nk}, (\mathbf{z}_{nk}^{(k)})^-) \geq F(\mathbf{z}_{nk}, (\mathbf{z}_{nk}^{(k)})^+)$, implying that latent vector $\mathbf{z}_{nk}$ should be more similar to another vector that shares the same covariate value, $(\mathbf{z}_{nk}^{(k)})^+$, than to a vector with a different covariate value, $(\mathbf{z}_{nk}^{(k)})^-$. Furthermore, the latent vector $\mathbf{z}_{n0}$ should show equal similarity to any other vectors regardless of their covariate values, whether $(\mathbf{z}_{n0}^{(k)})^-$ and $(\mathbf{z}_{n0}^{(k)})^+$, thus fulfilling the condition: $F(\mathbf{z}_{n0}, (\mathbf{z}_{n0}^{(k)})^-) = F(\mathbf{z}_{n0}, (\mathbf{z}_{n0}^{(k)})^+)$. This equality ensures $\mathbf{z}_{n0}$ remains unaffected by covariate-specific information, thereby providing a covariate-neutral representation of the cell's gene expression. Ultimately, the aim of *TarDis* is to produce a latent representation in which $\mathbf{z}_{nk}$ reflects the influence of its corresponding covariate $\mathbf{s}_{nk}$, while $\mathbf{z}_{n0}$ offers a covariate-neutral representation of the cell's gene expression profile, unaffected by any covariate-specific variations.

2.1. VAE Skeleton

TarDis builds upon a variational autoencoder (VAE) to construct a high-fidelity generative model that underpins our disentanglement objectives. The VAE component optimization guided by the Evidence Lower Bound (ELBO), a surrogate for the intractable marginal log-likelihood as shown in Equation 1 (Kingma & Welling, 2013). Here, the covariates, $\mathbf{s}_n$, are pivotal for capturing factors that might influence the observed data, such as batch effects. *TarDis* incorporates the target covariates as $\mathbf{s}_n$, and also allows inclusion of non-target covariates, providing flexibility in managing different types of data impacts. The first term of $\mathcal{L}_{\text{VAE}}$ represents the reconstruction loss, $\mathcal{L}_{\text{R}}$, which quantifies the expected negative log-likelihood of the observed data $\mathbf{x}_n$, conditioned on the latent variables, $\mathbf{z}_n$. The reconstruction loss is formally expressed using the negative binomial (NB) distribution, ideal for capturing the count variability inherent in data types like single-cell genomics (Equation 2). In this equation, Γ denotes the gamma function, $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$ refer to the mean and inverse dispersion parameters of the negative binomial distribution, respectively (Inecik & Theis, 2023). The second term measures the Kullback-Leibler divergence (KL), $\mathcal{L}_{\text{KL}}$, penalizing deviations of the learned posterior distribution $q_\phi(\mathbf{z}_n \mid \mathbf{x}_n, \mathbf{s}_n)$ from the prior distribution $p(\mathbf{z}_n)$. In Equation 3, the approximate posterior distribution is assumed to be Gaussian distribution with mean $\boldsymbol{\mu}_n$ and diagonal covariance matrix $\boldsymbol{\Sigma}_n$, and the prior distribution $p(\mathbf{z}_n)$ is typically a standard normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ where $\boldsymbol{I}$ is the identity matrix in $\mathbb{R}^{|\mathbf{z}_n| \times |\mathbf{z}_n|}$.

$$\mathcal{L}_{\text{VAE}}(\theta, \phi; \mathbf{x}_n, \mathbf{s}_n) = \big[- \mathbb{E}_{q_\phi(\mathbf{z}_n|\mathbf{x}_n,\mathbf{s}_n)} [\log p_\theta(\mathbf{x}_n|\mathbf{z}_n)] + D_{\text{KL}}(q_\phi(\mathbf{z}_n \mid \mathbf{x}_n, \mathbf{s}_n) \parallel p(\mathbf{z}_n)) \big] \tag{1}$$

$$\mathcal{L}_{\text{R}} = \frac{\Gamma(\mathbf{x}_n + \boldsymbol{\theta}_n)}{\Gamma(\mathbf{x}_n + 1)\Gamma(\boldsymbol{\theta}_n)} \left(\frac{\boldsymbol{\theta}_n}{\boldsymbol{\theta}_n + \boldsymbol{\mu}_n}\right)^{\boldsymbol{\theta}_n} \left(\frac{\boldsymbol{\mu}_n}{\boldsymbol{\theta}_n + \boldsymbol{\mu}_n}\right)^{\mathbf{x}_n} \tag{2}$$

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}(\mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) \parallel \mathcal{N}(\mathbf{0}, \boldsymbol{I})) \tag{3}$$

2.2. Auxiliary Loss

In *TarDis* model training, the VAE optimization is intertwined with the novel auxiliary loss component introduced, $\mathcal{L}_{\text{C}}$, to construct $\mathbf{z}_n = (\mathbf{z}_{n0}, [\mathbf{z}_{nk}]_{k \in J_k})$ with $\mathbf{z}_{nk} \sim \mathcal{N}(\boldsymbol{\mu}_{nk}, \boldsymbol{\Sigma}_{nk})$. The overall loss function of *TarDis* integrates these components through a weighted sum, controlled by hyperparameters λ_{C}, λ_{KL}, and λ_{R} (Equation 8). Specifically, $\mathcal{L}_{\text{C}}$ is a composite loss function that incorporates four distinct loss components for each covariate. For each target covariate $\mathbf{s}_{nk}$, the loss function, $\mathcal{L}_{\text{C}}^{(k)}$, includes $(N_{\mathcal{L}}^{(k)})^+$ positive and $(N_{\mathcal{L}}^{(k)})^-$ negative loss terms. Similarly, for the covariate-free representation $\mathbf{z}_{n0}$, it includes $(N_{\mathcal{L}}^{(k_0)})^+$ positive and $(N_{\mathcal{L}}^{(k_0)})^-$ negative terms. The losses for positive pairs and negative pairs given in Equations 4 and 5, respectively. Here, the λ values are hyperparameters that determine the weight of each loss component, while the $\mathcal{L}$ loss functions encompass metrics such as KL divergence and MSE[4]. Thus, the overall covariate loss, $\mathcal{L}_{\text{C}}^{(k)}$, is computed as the sum of these two pair losses, as specified in Equation 6. By aggregating these individual covariate losses, the total auxiliary loss, $\mathcal{L}_{\text{C}}$, is expressed in Equation 7.

The configuration of $\mathcal{L}_{\text{C}}^{(k)}$ is meticulously designed to meet several critical objectives within the *TarDis* framework. First, by minimizing the distance between $(\mathbf{z}_{nk}^{(k)})^+$ and $\mathbf{z}_{nk}$, the model ensures that the latent representations of positive examples closely align with their corresponding covariate within respective latent subset, accurately reflecting specific characteristics. In contrast, it maximizes the distance between $(\mathbf{z}_{nk}^{(k)})^-$ and $\mathbf{z}_{nk}$, thereby promoting clear differentiation in the latent representations of negative examples and enhancing the distinction between different covariates. Additionally, the model strategy involves maximizing the distance between $(\mathbf{z}_{n0}^{(k)})^+$ and $\mathbf{z}_{n0}$, while minimizing the distance between $(\mathbf{z}_{n0}^{(k)})^-$ and $\mathbf{z}_{n0}$. This approach ensures that $\mathbf{z}_{n0}$ remains free from covariate-specific influences, maintaining its role as a covariate-neutral representation. These

operations collectively ensure that covariate information is precisely captured in the respective targeted latent subsets, $\mathbf{z}_{nk}$, and effectively isolated from $\mathbf{z}_{n0}$.

$$(\mathcal{L}_{\mathrm{C}}^{(k)})^{+}(\phi;\mathbf{x}_n,\mathbf{s}_n) = \Bigg[\sum_{i=1}^{(N_{\mathcal{L}}^{(k)})^{+}} \frac{(\lambda_{\mathrm{C}}^{(k)})_i^{+}(\mathcal{L}_{\mathrm{C}}^{(k)})_i^{+}(\phi;\mathbf{x}_n,\mathbf{s}_n)}{(N_{\mathcal{L}}^{(k)})^{+}} + \sum_{i=1}^{(N_{\mathcal{L}}^{(k_0)})^{+}} \frac{(\lambda_{\mathrm{C}}^{(k_0)})_i^{+}(\mathcal{L}_{\mathrm{C}}^{(k_0)})_i^{+}(\phi;\mathbf{x}_n,\mathbf{s}_n)}{(N_{\mathcal{L}}^{(k_0)})^{+}}\Bigg] \tag{4}$$

$$(\mathcal{L}_{\mathrm{C}}^{(k)})^{-}(\phi;\mathbf{x}_n,\mathbf{s}_n) = \Bigg[\sum_{i=1}^{(N_{\mathcal{L}}^{(k)})^{-}} \frac{(\lambda_{\mathrm{C}}^{(k)})_i^{-}(\mathcal{L}_{\mathrm{C}}^{(k)})_i^{-}(\phi;\mathbf{x}_n,\mathbf{s}_n)}{(N_{\mathcal{L}}^{(k)})^{-}} + \sum_{i=1}^{(N_{\mathcal{L}}^{(k_0)})^{-}} \frac{(\lambda_{\mathrm{C}}^{(k_0)})_i^{-}(\mathcal{L}_{\mathrm{C}}^{(k_0)})_i^{-}(\phi;\mathbf{x}_n,\mathbf{s}_n)}{(N_{\mathcal{L}}^{(k_0)})^{-}}\Bigg] \tag{5}$$

$$\mathcal{L}_{\mathrm{C}}^{(k)}(\phi;\mathbf{x}_n,\mathbf{s}_n) = (\mathcal{L}_{\mathrm{C}}^{(k)})^{+}(\phi;\mathbf{x}_n,\mathbf{s}_n) + (\mathcal{L}_{\mathrm{C}}^{(k)})^{-}(\phi;\mathbf{x}_n,\mathbf{s}_n) \tag{6}$$

$$\mathcal{L}_{\mathrm{C}}(\phi;\mathbf{x}_n,\mathbf{s}_n) = \frac{1}{|J_K|}\sum_{k=1}^{|J_K|}\mathcal{L}_{\mathrm{C}}^{(k)}(\phi;\mathbf{x}_n,\mathbf{s}_n) \tag{7}$$

$$\mathcal{L}_{\text{TarDis}} = \lambda_{\mathrm{C}}\mathcal{L}_{\mathrm{C}} + \lambda_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}} + \lambda_{\mathrm{R}}\mathcal{L}_{\mathrm{R}} \tag{8}$$

Although the theoretical framework primarily employs KL divergence as the loss metric, the principle is also applicable to various losses between anchor points and negative or positive samples with minor adjustments [4]. The optimization of various hyperparameters, including the individual loss weights, is conducted once and uniformly applied across all experiments, unless explicitly stated otherwise [5]. The experiments and benchmarking processes utilize a diverse array of datasets to ensure comprehensive testing and validation of the model. Each dataset is selected to represent different types and scales of data challenges [6]. Various evaluation metrics are used to assess the model's performance, with a full discussion provided in Appendix H [7]. The assumptions behind the theoretical framework are discussed in Appendix E[8].

3. Results

3.1. *TarDis* achieves robust disentanglement of covariates into isolated latent spaces

We assessed the *TarDis* model's ability to disentangle covariates using the *Afriat* single-cell genomics dataset, which includes three distinct covariates: age, zone status, and time (Appendix C.1). Experiments were conducted with two methodologies: disentangling all covariates simultaneously, *TarDis*$_{\text{multiple}}$[9], and disentangling each covariate individually followed by concatenating the reserved latent spaces, *TarDis*$_{\text{single}}$. The disentanglement performance was benchmarked using the maximum mutual information gap (maxMIG), as detailed in Figure 1a, demonstrating that both configurations of *TarDis* surpassed existing models (Appendix A) and achieved nearly 0.9 maxMIG scores on validation sets (Shamsaie et al., 2024). These results underscore the efficacy of *TarDis* in handling multiple covariates simultaneously without compromising disentanglement quality. Further analysis using the mutual information (MI) metric reveals minimal differences in the preservation of information within the unreserved and reserved latent spaces between the two training strategies, indicating the model's effective scalability for disentanglement tasks (Figure 1b). Notably, for all subsequent experiments detailed in this paper, we have exclusively employed the multiple-covariate disentanglement approach.

An ablation study was performed to evaluate the model's robustness against feature reduction, where varying percentages of input features were systematically removed. Results in Figure 1c show that *TarDis* maintained high maxMIG and R^2 reconstruction scores, above 0.65 and 0.94 respectively, affirming its resilience to input variability. Additionally, modifying the auxiliary loss weight, λ_{C}, systematically influenced the clustering quality and disentanglement accuracy, as indicated by the increased maxMIG score and mean centroid distance with higher λ_{C} values (Figure 1d and Supplementary Figure 7). Moreover, the silhouette scores, calculated on the unreserved latent space $\mathbf{z}_{n0}$ using cell type annotations as the labels, provided empirical evidence that effective disentanglement correlates with enhanced biological signal representation, as further investigated in Results 3.2. Overall, these results not only validate the robustness of *TarDis* in disentangling complex covariate structures but also highlight its utility in preserving essential biological variations, pivotal for advancing single-cell genomic data analysis.

3.2. *TarDis* achieves superior performance in single-cell genomics data integration

To probe the efficacy of invariant representation learning, we turned our attention to the *Suo* dataset, a massive single-cell genomics dataset capturing human embryonic development. This dataset includes about 850k cells from various organs and time points, using multiple methods, instruments, samples, and platforms, as well as a wide range of cell types (Appendix C.2). Its complexity makes it an ideal testbed for evaluating model performance in integrating intricate

[4] See Appendix F for a discussion on the loss function options.
[5] See Appendix G for experiment hyperparameters and settings.
[6] See Appendix C for a description of the datasets used.
[7] See Appendix H for the evaluation metrics employed.
[8] See Appendix E for a discussion of the model assumptions.
[9] See Supplementary Figure 6 for UMAP visualizations of reserved latent space representations.

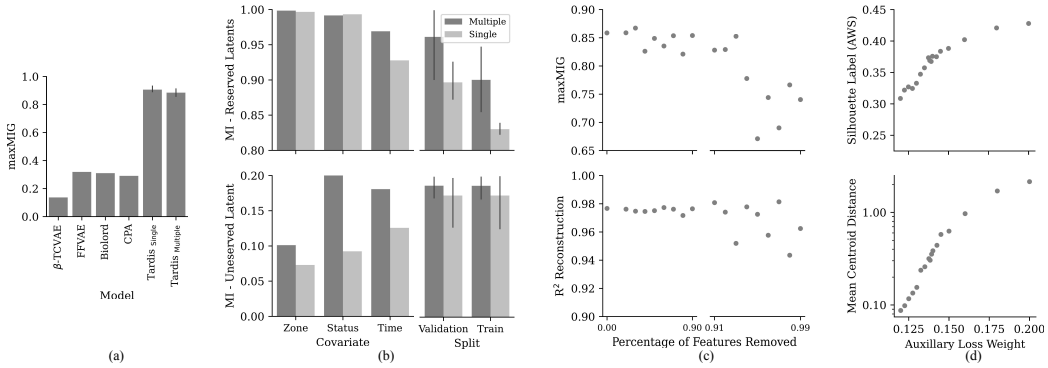


Figure 1. (a) Comparison of disentanglement performance using maxMIG, showing that *TarDis* variants outperform existing models. (b) MI in the reserved, $\mathbf{z}_{nk}$, and unreserved, $\mathbf{z}_{n0}$, latent spaces for *TarDis* under multiple and single covariate training conditions across various covariates and data splits. (c) Relationship between the percentage of input features removed and the corresponding maxMIG and R^2 reconstruction scores, indicating robustness to feature removal. (d) Impact of auxiliary loss weight (λ_C) on mean centroid distance in reserved latents, $\mathbf{z}_{nk}$, and average silhouette width (ASW) scores at the unreserved latent, $\mathbf{z}_{n0}$.

datasets. We assessed the data integration quality using the scIB package metrics (Luecken et al., 2022), which are recognized benchmarks in the single-cell genomics community for evaluating the balance between biological signal preservation and batch effect mitigation (Appendix H). This balance is crucial as inadequate correction can lead to data clustering by batch, obscuring true biological variance, while over-correction may suppress biological signals, reducing the biological relevance of the outcomes.

In our experimental setup, we tested two configurations of the *TarDis* model. *TarDis*-1 focuses on covariates typically considered as batch keys in single-cell data integration tasks, such as library platform, donor, sample status, and instruments. *TarDis*-2 extends this disentanglement to additional covariates including sex, age, and notably, organ. The comparative results, detailed in Table 1, show that *TarDis*, particularly *TarDis*-2, outperforms state-of-the-art models[10] and maintains an optimal balance between biological conservation and batch correction. By effectively disentangling various spurious correlations from invariant biological signals, *TarDis* has demonstrated its robust capability to manage the complexities inherent in vast and heterogeneous datasets.

3.3. *TarDis* generates ordered latent representation for continuous covariates

In addressing the challenge of learning the representation of disentangled *continuous* covariates, *TarDis* provides a

[10]All models were trained under configurations that aimed to closely mirror the training of *TarDis* models given in Appendix G, ensuring consistency in architectural choices and the selection of analogous hyperparameters where applicable.

Table 1. Benchmarking data integration performance by scIB package (Luecken et al., 2022) metrics, organized into biological signal conservation and batch correction categories (Appendix H). Quantification employed a comprehensive set of metrics, with aggregate scores derived according to scIB standards. Cell-type annotations are incorporated in the metrics where *labels* are necessary. Covariates such as library platform, donor, sample status, and instrument are used as *batch keys* when required.

	Metric	PCA	Harmony	scVI	scANVI	inVAE	*TarDis*-1	*TarDis*-2
Bio conservation	Isolated Labels	0.610	0.563	0.638	0.774	0.798	0.662	0.767
	K-means NMI	0.691	0.620	0.649	0.792	0.651	0.634	0.713
	K-means ARI	0.226	0.182	0.209	0.360	0.191	0.185	0.228
	Silhouette Label (AWS)	0.504	0.482	0.496	0.576	0.508	0.497	0.508
	Cell-type LISI (cLISI)	0.999	0.997	0.999	1.000	0.999	0.998	0.999
Batch correction	Silhouette Batch	0.851	0.862	0.867	0.861	0.840	0.903	0.896
	Integration LISI (iLISI)	0.057	0.100	0.098	0.093	0.040	0.094	0.080
	kBET per Label	0.309	0.475	0.487	0.526	0.194	0.448	0.430
	Graph Connectivity	0.793	0.671	0.866	0.912	0.836	0.866	0.879
	PCR Comparison	0.000	0.350	0.699	0.222	0.000	0.931	0.850
Aggregate score	Bio conservation	0.606	0.569	0.598	**0.701**	0.629	0.595	0.643
	Batch correction	0.402	0.492	0.603	0.523	0.382	**0.648**	0.627
	Total	0.524	0.538	0.600	0.629	0.530	0.616	**0.637**

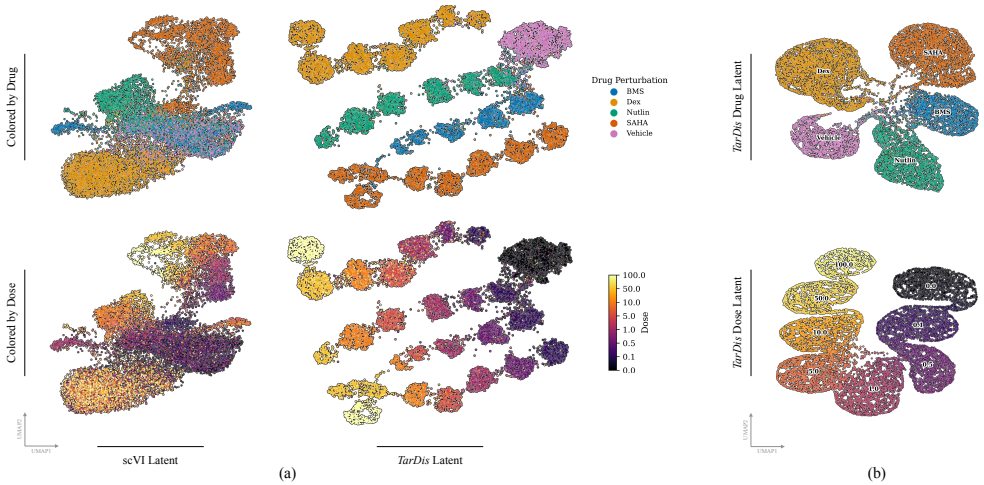


Figure 2. UMAP visualization (McInnes et al., 2018) of *TarDis* latent space representations from the *Sciplex* dataset (a) Comparing the performance of scVI and *TarDis* models in capturing drug responses and dosage effects. The upper row displays clusters differentiated by drug types, while the bottom row illustrates the ordered representation of dosage, showcasing the ability of *TarDis* to structurally organize cellular responses across different drug concentrations. (b) *TarDis* model training generates three distinct latent spaces: unreserved, dose, and drug. Displayed UMAPs are the dose and drug latent subspaces, demonstrating structured separation and ordered representation.

solution that captures data variations without reducing them to mere categorical approximations. Continuous covariates such as age or treatment dosage are critical for understanding gradients in biological processes, cellular behavior, and disease progression. To manage the subtleties associated with these variables, *TarDis* employs a distance-based loss function for each auxiliary loss component. The model employs negative pair losses weighted by the distance between the values of the continuous covariates, omitting positive pair losses due to the continuous nature of the covariate, which results in generating an ordered and interpretable latent space (Figure 2).

We here focused on two primary continuous covariates, age and drug dosage, which present distinct challenges due to their variability and significant impact on cellular phenotypes. We employed two datasets to evaluate the effectiveness of *TarDis* in producing ordered latent representations of these covariates. The first dataset, named *Sciplex* (Appendix C.5), involves drug perturbation experiments and helps in analyzing the structured response of cells to varying drug dosages. The second dataset, referred to as *Braun* (Appendix C.3), comprises 1.6 million cells from human embryonic brain development, providing a complex scenario for assessing the impact of time as a continuous variable. Through *TarDis*, we managed to produce ordered latent representations of these covariates within isolated latent subsets while concurrently disentangling other variables such as the type of library platform, donor characteristics, sample status, instrumentation used, and tissue types (Figure 2, 3).

This representation has enabled previously unfeasible hypothesis-driven biological analyses. For example, *TarDis* allows for the exploration of organ-specific developmental gene expression patterns for specific cell types, an analysis that previously wasn't optimal with non-batch-corrected input spaces. Unlike existing models such as scVI and scANVI, which address batch effects but often fail to retain essential biological information like age or organ specifics —either being overly corrected by batch keys or inadequately accounted for (Lopez et al., 2018; Xu et al., 2021)— *TarDis* allows researchers to isolate cells from two different organs using the organ-specific latent subset and, for a given cell type, compare expression patterns across developmental stages in a massive multi-organ developmental single-cell dataset. This analysis benefits from a batch-corrected latent space, thanks to a set of other latent subsets that disentangle batch effects. In Figure 3 upper right, *TarDis* enabled to identify genes including `EGR2-3-4`, `KLF2-4`, `RTL1`, `SPRY4-AS1`, and `FOSB`, that decrease in expression through embryonic development of human *forebrain neurons* within the *Braun* dataset, which were shown to be associated with brain development, aging, and diseases including Down syndrome and bipolar disorder (Manning et al., 2019; Chou et al., 2022; Kitazawa et al., 2021; Yin et al., 2015; Palmer et al., 2021; Poirier et al., 2007). In a parallel experiment using the *Sciplex* perturbation dataset, *TarDis* effectively disentangled the influences of drug type and dosage (Figure 2, 3). Using the data points corresponding to *Nutlin* cluster in drug latent, we analyzed how gene expression responds to increasing doses. As shown in Figure 3 bottom right, this approach allowed us to pinpoint the expression patterns of genes such as `TP53I3`, `CDKN1A`, `GDF15`, `MDM2`, `FDXR`, and `NUPR1`, which are notably responsive to escalating doses of *Nutlin* (Voltan et al., 2014; Huang & Vassilev, 2009).

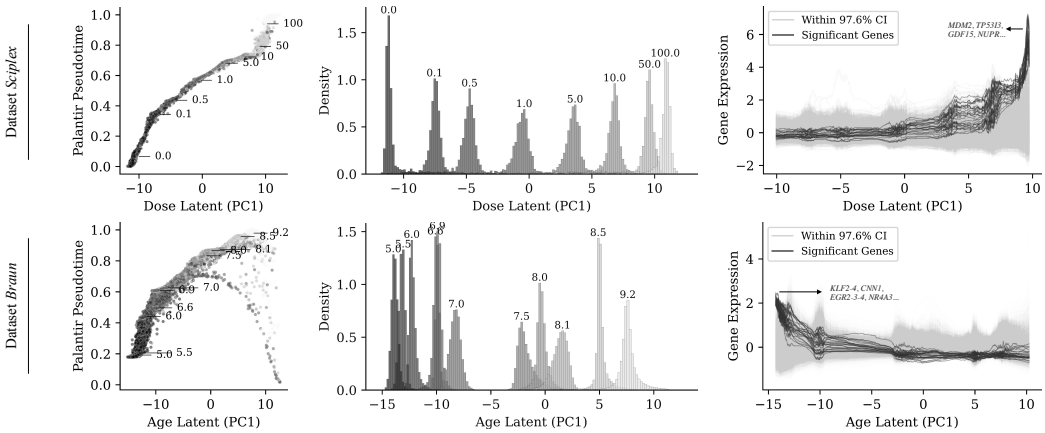


Figure 3. Ordered latent spaces for dose and age (post-conception week) in the *Sciplex* and *Braun* datasets, respectively. (left) Principal Component 1 (PC1) of the continuous covariate latent space plotted against Palantir pseudotime (Setty et al., 2019), which uses a k-nearest neighbor graph to infer cell pseudotime trajectories. (middle) Density distribution of the continuous covariate in the respective latent subset, illustrating ordered peaks corresponding to varying levels of the covariate. (Right) Differential gene expression profiles plotted against the continuous covariate latent space, identifying genes that show variation in expression levels associated with changes in the covariate, indicative of underlying cellular processes. Gene expression patterns are highlighted with (upper right) increasing doses of *Nutlin* and (bottom right) through human embryonic developmental stages of *forebrain neuron*.

3.4. *TarDis* predicts counterfactual gene expressions accurately under OOD conditions

The capacity of predictive models to generate accurate gene expressions under OOD conditions is pivotal for extrapolating research findings to new or novel environments. In evaluating this capacity, *TarDis* was systematically tested using two distinct datasets to gauge its effectiveness in predicting counterfactual gene expressions. Using the *Afriat* dataset, previously introduced, multiple models were trained, each excluding a different combination of three covariates to create respective OOD sets. Additionally, the *Miller* dataset, which comprises samples from human developmental embryonic lung, was utilized to disentangle the effects of age and donor covariates (Appendix C.4). Similar to the *Afriat* dataset, combinations of two covariates were systematically omitted during training to simulate various OOD conditions.

TarDis demonstrated superior performance in predicting gene expressions under OOD conditions, outperforming CPA[10], another model that *concurrently* disentangles multiple covariates, in both the *Afriat* and *Miller* datasets. In the *Afriat* dataset, *TarDis* achieved notably higher R^2 reconstruction scores, showcasing its strong capability for accurate reconstruction under varied and unseen conditions (Appendix H.12). In the *Miller* dataset, the challenge intensified with the evaluation focusing on differentially expressed genes, DEGs, (Appendix H.13). *TarDis* excelled, achieving significantly better OOD predictions for DEGs compared to CPA. These results, shown in Figure 4, affirm the utility of *TarDis* not only in disentangling complex covariate interactions within datasets but also in its capability to generalize across novel, unseen domains, key to advancing the precision and reliability of predictive models in single-cell genomics.

3.5. *TarDis* produces interpretable latent representations of disentangled covariates

In exploring the capabilities of *TarDis* to yield interpretable latent representations, we utilized the *Norman* dataset, a comprehensive collection comprising 108k cells subjected to single or combinatorial gene perturbations (Appendix C.6). This dataset is particularly challenging due to the diversity and complexity of its perturbations, with a total of 284 distinct perturbation conditions included in

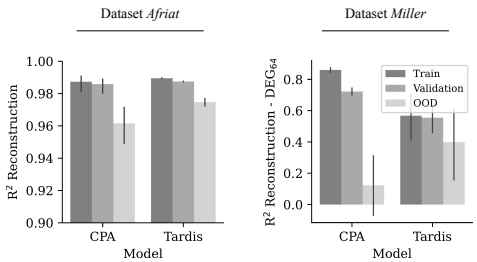
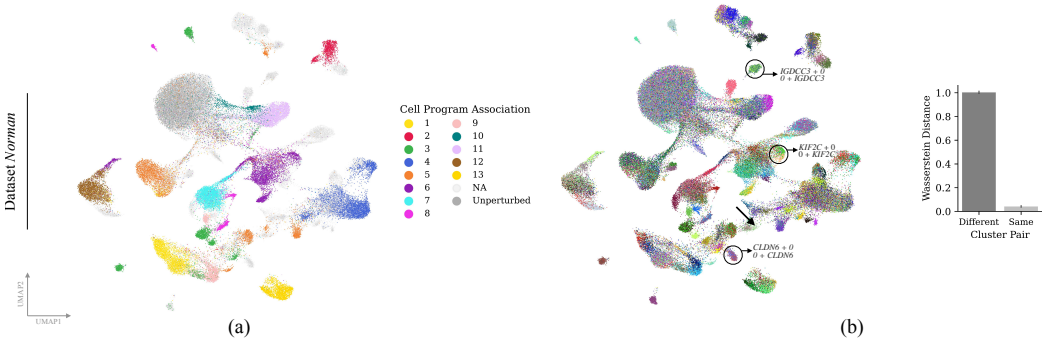


Figure 4. Performance comparison of *TarDis* and CPA in predicting counterfactual gene expressions under out-of-distribution conditions using the *Afriat* and *Miller* datasets. R^2 scores for reconstructed gene expressions and differentially expressed genes (DEG) across varying unseen covariate combinations highlight *TarDis*'s superior predictive capabilities.



(a) (b)

Figure 5. UMAP visualization of the *TarDis* perturbation latent space derived from the *Norman* dataset: (a) Clusters corresponding to sets of perturbations associated with similar cell programs as identified in the original publication (Norman et al., 2019), demonstrating the model's ability to capture underlying biological patterns. (b) UMAP visualization of *TarDis* latent space, colored by 270 perturbations. Representative clusters are highlighted, illustrating the model's capability to align identical perturbations accurately despite nominal labeling differences, thus confirming label reconciliation. Wasserstein distances are computed to quantitatively confirm the close, often overlapping, clustering of identical perturbations (Vallender, 1974).

this analysis. In this experiment, the inference model in *TarDis* relied solely on input features without the introduction of covariate information, $\mathbf{s}_n$. This approach ensured that the learning process was purely driven by the data's inherent structure rather than external annotations. Our results indicate that *TarDis* effectively disentangles these perturbations, with each perturbation distinctly isolated in the latent space. Significantly, perturbations that share a common cellular program, as identified in the original publication of the dataset (Norman et al., 2019), were found to cluster closely. The results support *TarDis* ability to capture interpretable and biologically meaningful patterns, as the clustering is not random qualitatively but reflects the underlying biological relationships (Figure 5a).

A particularly rigorous test of the model's interpretability involved the re-labeling of certain perturbations in the dataset. Specifically, the labels were altered to appear as two distinct entities: *'X+0'* and *'0+X'*, despite originating from the same perturbation. This was designed to test whether *TarDis* could recognize and reconcile these as identical despite their nominal differences. The results were in line with our expectations: *TarDis* successfully overlapped these perturbations in the latent space, affirming its capability to generate biologically coherent and interpretable latent representations, even under challenging conditions (Figure 5b). This analysis not only confirms the robustness of *TarDis*'s disentanglement capabilities but also highlights its potential in generating actionable insights from complex genomic data, where interpretability is crucial for meaningful biological inference.

4. Conclusion

In this study, we presented *TarDis*, a novel deep generative model designed for the targeted disentanglement of covariates in complex multi-domain and multi-condition datasets, particularly focusing on the challenges presented by single-cell genomics data. Our approach leverages a series of covariate-specific loss functions to facilitate robust disentanglement and invariant representation of both continuous and categorical variables, thus enhancing data integration capabilities and enabling more insightful biological analyses. Through rigorous benchmarking against existing models and diverse datasets, *TarDis* has demonstrated superior performance not only in its capacity to disentangle complex covariate structures but also in maintaining essential biological signals crucial for accurate data interpretation and analysis, and generating robust predictions under out-of-distribution conditions. Moreover, *TarDis*'s ability to generate ordered latent representations of continuous covariates significantly enhances differential analyses across varying conditions. The model perform robustly in generating interpretable and biologically meaningful latent representations, which could empower researchers to conduct advanced hypothesis-driven research, potentially unveiling novel insights and therapeutic targets.

TarDis establishes a robust approach for exploring complex biological questions, offering researchers unprecedented clarity in dissecting the nuanced interactions between diverse covariates. This capability is instrumental in advancing personalized medicine, supporting the development of customized therapeutic strategies grounded in a profound understanding of individual responses to different treatments. Considering the expansion of *TarDis* applications beyond genomics, for instance into neuromarketing using EEG event-related potentials (ERP) data, it becomes crucial to acknowledge that modifications to the model may be necessary to accommodate different types of data. We are actively investigating these potential applications, aiming to extend the reach and impact of *TarDis* across various scientific and applied fields[11].

[11] Refer to Appendix D for a discussion regarding the limitations.

References

Afriat, A., Zuzarte-Luís, V., Bahar Halpern, K., Buchauer, L., Marques, S., Chora, A. F., Lahree, A., Amit, I., Mota, M. M., and Itzkovitz, S. A spatiotemporally resolved single-cell atlas of the plasmodium liver stage. *Nature*, 611(7936):563–569, 2022.

Ahuja, K., Wang, J., Dhurandhar, A., Shanmugam, K., and Varshney, K. R. Empirical or invariant risk minimization? a sample complexity perspective. *arXiv preprint arXiv:2010.16412*, 2020.

Ahuja, K., Caballero, E., Zhang, D., Gagnon-Audet, J.-C., Bengio, Y., Mitliagkas, I., and Rish, I. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:3438–3450, 2021.

Aliee, H., Kapl, F., Hediyeh-Zadeh, S., and Theis, F. J. Conditionally invariant representation learning for disentangling cellular heterogeneity. *arXiv preprint arXiv:2307.00558*, 2023.

Andéol, L., Kawakami, Y., Wada, Y., Kanamori, T., Müller, K.-R., and Montavon, G. Learning domain invariant representations by joint wasserstein distance minimization. *Neural Networks*, 167:233–243, 2023.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Azzam, M., Gnanha, A. T., Wong, H.-S., and Wu, S. Adversarially constrained interpolation for unsupervised domain adaptation. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 2375–2381. IEEE, 2021.

Baker, D. N., Dyjack, N., Braverman, V., Hicks, S. C., and Langmead, B. Fast and memory-efficient scrna-seq k-means clustering with various distances. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '21. ACM, August 2021.

Bergen, V., Lange, M., Peidli, S., Wolf, F. A., and Theis, F. J. Generalizing rna velocity to transient cell states through dynamical modeling. *Nature biotechnology*, 38 (12):1408–1414, 2020.

Beucler, T., Gentine, P., Yuval, J., Gupta, A., Peng, L., Lin, J., Yu, S., Rasp, S., Ahmed, F., O'Gorman, P. A., et al. Climate-invariant machine learning. *Science Advances*, 10(6):eadj7250, 2024.

Biology, C. S.-C., Abdulla, S., Aevermann, B., Assis, P., Badajoz, S., Bell, S. M., Bezzi, E., Cakir, B., Chaffer, J., Chambers, S., et al. Cz cellxgene discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *bioRxiv*, pp. 2023–10, 2023.

Braun, E., Danan-Gotthold, M., Borm, L. E., Lee, K. W., Vinsland, E., Lönnerberg, P., Hu, L., Li, X., He, X., Andrusivová, Ž., et al. Comprehensive cell atlas of the first-trimester developing human brain. *Science*, 382(6667): eadf1226, 2023.

Bunne, C., Stark, S. G., Gut, G., Del Castillo, J. S., Levesque, M., Lehmann, K.-V., Pelkmans, L., Krause, A., and Rätsch, G. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20(11):1759–1768, 2023.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420, 2018.

Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A., and Theis, F. J. A test metric for assessing single-cell rna-seq batch correction. *Nature Methods*, 16(1):43–49, December 2018. ISSN 1548-7105.

Cao, D., Zhou, S., Liu, H., Liu, J., and Zang, H. Signal censoring and fusing with system-level communication constraints in multistatic radar: a j-divergence and bhattacharyya distance-based approach. *IET Radar, Sonar & Navigation*, 11(12):1802–1814, 2017.

Cao, Z., Yu, H., Yang, H., and Sano, A. Pirl: participant-invariant representation learning for healthcare using maximum mean discrepancy and triplet loss. *arXiv preprint arXiv:2302.09126*, 2023.

Chen, R. T., Li, X., Grosse, R. B., and Duvenaud, D. K. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

Chou, M.-Y., Hu, M.-C., Chen, P.-Y., Hsu, C.-L., Lin, T.-Y., Tan, M.-J., Lee, C.-Y., Kuo, M.-F., Huang, P.-H., Wu, V.-C., et al. Rtl1/peg11 imprinted in human and mouse brain mediates anxiety-like and social behaviors and regulates neuronal excitability in the locus coeruleus. *Human Molecular Genetics*, 31(18):3161–3180, 2022.

De Donno, C., Hediyeh-Zadeh, S., Moinfar, A. A., Wagenstetter, M., Zappia, L., Lotfollahi, M., and Theis, F. J. Population-level integration of single-cell datasets enables multi-scale analysis across samples. *Nature Methods*, 20(11):1683–1692, 2023.

Duncan, T. E. On the calculation of mutual information. *SIAM Journal on Applied Mathematics*, 19(1):215–220, 1970.

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., and Theis, F. J. Single-cell rna-seq denoising using a deep count autoencoder. *Nature communications*, 10(1):390, 2019.

Goel, K., Gu, A., Li, Y., and Ré, C. Model patching: Closing the subgroup performance gap with data augmentation. *arXiv preprint arXiv:2008.06775*, 2020.

Guo, S., Tóth, V., Schölkopf, B., and Huszár, F. Causal de finetti: On the identification of invariant causal structure in exchangeable data. *Advances in Neural Information Processing Systems*, 36, 2024.

Haghverdi, L., Lun, A. T. L., Morgan, M. D., and Marioni, J. C. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427, April 2018. ISSN 1546-1696. doi: 10.1038/nbt.4091.

Hajihassnai, O., Ardakanian, O., and Khazaei, H. Obscurenet: Learning attribute-invariant latent representation for anonymizing sensor data. In *Proceedings of the international conference on internet-of-things design and implementation*, pp. 40–52, 2021.

Haniffa, M., Taylor, D., Linnarsson, S., Aronow, B. J., Bader, G. D., Barker, R. A., Camara, P. G., Camp, J. G., Chédotal, A., Copp, A., et al. A roadmap for the human developmental cell atlas. *Nature*, 597(7875):196–205, 2021.

Hetzel, L., Boehm, S., Kilbertus, N., Günnemann, S., Theis, F., et al. Predicting cellular responses to novel drug perturbations at a single-cell resolution. *Advances in Neural Information Processing Systems*, 35:26711–26722, 2022.

Heumos, L., Schaar, A. C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücken, M. D., Strobl, D. C., Henao, J., Curion, F., et al. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, pp. 1–23, 2023.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3, 2017.

Hrovatin, K., Moinfar, A. A., Lapuerta, A. T., Zappia, L., Lengerich, B., Kellis, M., and Theis, F. J. Integrating single-cell rna-seq datasets with substantial batch effects. *bioRxiv*, 2023.

Huang, B. and Vassilev, L. T. Reduced transcriptional activity in the p53 pathway of senescent cells revealed by the mdm2 antagonist nutlin-3. *Aging (Albany NY)*, 1(10): 845, 2009.

Huang, H., Chen, M., and Qiao, X. Generative learning for financial time series with irregular and scale-invariant patterns. In *The Twelfth International Conference on Learning Representations*, 2023.

Hubert, L. and Arabie, P. Comparing partitions. *Journal of Classification*, 2(1):193–218, December 1985a. ISSN 1432-1343.

Hubert, L. J. and Arabie, P. Comparing partitions. *Journal of Classification*, 2:193–218, 1985b.

Inecik, K. and Theis, F. J. scare: Attribution regularization for single cell representation learning. *bioRxiv*, pp. 2023–07, 2023.

Inecik, K., Uhlmann, A., Lotfollahi, M., and Theis, F. Multicpa: Multimodal compositional perturbation autoencoder. *bioRxiv*, pp. 2022–07, 2022.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217. PMLR, 2020.

Kim, H. and Mnih, A. Disentangling by factorising. In *International conference on machine learning*, pp. 2649–2658. PMLR, 2018.

Kim, M., Wang, Y., Sahu, P., and Pavlovic, V. Relevance factor vae: Learning and identifying disentangled factors. *arXiv preprint arXiv:1902.01568*, 2019.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kitazawa, M., Sutani, A., Kaneko-Ishino, T., and Ishino, F. The role of eutherian-specific rtl1 in the nervous system and its implications for the kagami-ogata and temple syndromes. *Genes to Cells*, 26(3):165–179, 2021.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

Kong, L., Xie, S., Yao, W., Zheng, Y., Chen, G., Stojanov, P., Akinwande, V., and Zhang, K. Partial disentanglement for domain adaptation. In *International conference on machine learning*, pp. 11455–11472. PMLR, 2022.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-r., and Raychaudhuri, S. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature Methods*, 16 (12):1289–1296, November 2019. ISSN 1548-7105.

Kraskov, A., Stögbauer, H., and Grassberger, P. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

Kumar, A., Sattigeri, P., and Balakrishnan, A. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017.

Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L., and Ranzato, M. Fader networks: Manipulating images by sliding attributes, 2018.

Lin, E., Mukherjee, S., and Kannan, S. A deep adversarial variational autoencoder model for dimensionality reduction in single-cell rna sequencing analysis. *BMC bioinformatics*, 21:1–11, 2020.

Liu, R., Qian, K., He, X., and Li, H. Integration of scrna-seq data by disentangled representation learning with condition domain adaptation. *BMC bioinformatics*, 25 (1):116, 2024.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

Lotfollahi, M., Klimovskaia Susmelj, A., De Donno, C., Hetzel, L., Ji, Y., Ibarra, I. L., Srivatsan, S. R., Naghipourfar, M., Daza, R. M., Martin, B., et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular systems biology*, 19(6):e11517, 2023.

Lu, C., Wu, Y., Hernández-Lobato, J. M., and Schölkopf, B. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*, 2021a.

Lu, C., Wu, Y., Hernández-Lobato, J. M., and Schölkopf, B. Nonlinear invariant risk minimization: A causal approach. *arXiv preprint arXiv:2102.12353*, 2021b.

Luecken, M. D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Müller, M. F., Strobl, D. C., Zappia, L., Dugas, M., Colomé-Tatché, M., et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature methods*, 19(1):41–50, 2022.

Manning, C. E., Eagle, A. L., Kwiatkowski, C. C., Achargui, R., Woodworth, H., Potter, E., Ohnishi, Y., Leinninger, G. M., and Robison, A. Hippocampal subgranular zone fosb expression is critical for neurogenesis and learning. *Neuroscience*, 406:225–233, 2019.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Miller, A. J., Yu, Q., Czerwinski, M., Tsai, Y.-H., Conway, R. F., Wu, A., Holloway, E. M., Walker, T., Glass, I. A., Treutlein, B., et al. In vitro and in vivo development of the human airway at single-cell resolution. *Developmental cell*, 53(1):117–128, 2020.

Moon, K. R., Stanley, J. S., Burkhardt, D., van Dijk, D., Wolf, G., and Krishnaswamy, S. Manifold learning-based methods for analyzing single-cell rna-sequencing data. *Current Opinion in Systems Biology*, 7:36–46, February 2018. ISSN 2452-3100.

Muus, C., Luecken, M. D., Eraslan, G., Sikkema, L., Waghray, A., Heimberg, G., Kobayashi, Y., Vaishnav, E. D., Subramanian, A., Smillie, C., et al. Single-cell meta-analysis of sars-cov-2 entry genes across tissues and demographics. *Nature medicine*, 27(3):546–559, 2021.

Norman, T. M., Horlbeck, M. A., Replogle, J. M., Ge, A. Y., Xu, A., Jost, M., Gilbert, L. A., and Weissman, J. S. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019.

Oh, C., Won, H., So, J., Kim, T., Kim, Y., Choi, H., and Song, K. Learning fair representation via distributional contrastive disentanglement. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1295–1305, 2022.

Palmer, C. R., Liu, C. S., Romanow, W. J., Lee, M.-H., and Chun, J. Altered cell and rna isoform diversity in aging down syndrome brains. *Proceedings of the National Academy of Sciences*, 118(47):e2114326118, 2021.

Perez, R. K., Gordon, M. G., Subramaniam, M., Kim, M. C., Hartoularos, G. C., Targ, S., Sun, Y., Ogorodnikov, A., Bueno, R., Lu, A., et al. Single-cell rna-seq reveals cell type–specific molecular and genetic associations to lupus. *Science*, 376(6589):eabf1970, 2022.

Piran, Z., Cohen, N., Hoshen, Y., and Nitzan, M. Disentanglement of single-cell data with biolord. *Nature Biotechnology*, pp. 1–6, 2024.

Poirier, R., Cheval, H., Mailhes, C., Charnay, P., Davis, S., and Laroche, S. Paradoxical role of an egr transcription factor family member, egr2/krox20, in learning and memory. *Frontiers in behavioral neuroscience*, 1:163, 2007.

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., et al. The human cell atlas. *elife*, 6: e27041, 2017.

Rood, J. E., Maartens, A., Hupalowska, A., Teichmann, S. A., and Regev, A. Impact of the human cell atlas on medicine. *Nature medicine*, 28(12):2486–2496, 2022.

Rousseeuw, P. Rousseeuw, p.j.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. comput. appl. math. 20, 53-65. *Journal of Computational and Applied Mathematics*, 20:53–65, 11 1987a. doi: 10.1016/0377-0427(87)90125-7.

Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, November 1987b. ISSN 0377-0427.

Sepliarskaia, A., Kiseleva, J., and de Rijke, M. How to not measure disentanglement. *arXiv preprint arXiv:1910.05587*, 2019.

Setty, M., Kiseliovas, V., Levine, J., Gayoso, A., Mazutis, L., and Pe'Er, D. Characterization of cell fate probabilities in single-cell data with palantir. *Nature biotechnology*, 37(4):451–460, 2019.

Shamsaie, K., Megas, S., Asadollahzadeh, H., Teichmann, S. A., and Lotfollahi, M. Disentangling covariates to predict counterfactuals for single-cell data, 2024. URL https://openreview.net/forum?id=YeOUqnPVwM.

Shree, A., Pavan, M. K., and Zafar, H. scdreamer for atlas-level integration of single-cell datasets using deep generative model paired with adversarial classifier. *Nature Communications*, 14(1):7781, 2023.

Sikkema, L., Ramírez-Suástegui, C., Strobl, D. C., Gillett, T. E., Zappia, L., Madissoon, E., Markov, N. S., Zaragosi, L.-E., Ji, Y., Ansari, M., et al. An integrated cell atlas of the lung in health and disease. *Nature Medicine*, 29(6):1563–1577, 2023.

Silva, W. B., Freitas, C. C., Sant'Anna, S. J., and Frery, A. C. Classification of segments in polsar imagery by minimum stochastic distances between wishart distributions. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(3):1263–1273, 2013.

Sintini, L. and Kunze, L. Unsupervised and semi-supervised novelty detection using variational autoencoders in opportunistic science missions. In *British Machine Vision Conference*, 2020.

Sokolova, M. and Lapalme, G. A systematic analysis of performance measures for classification tasks. *Information Processing; Management*, 45(4):427–437, July 2009. ISSN 0306-4573. doi: 10.1016/j.ipm.2009.03.002.

Srivatsan, S. R., McFaline-Figueroa, J. L., Ramani, V., Saunders, L., Cao, J., Packer, J., Pliner, H. A., Jackson, D. L., Daza, R. M., Christiansen, L., et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020.

Sturma, N., Squires, C., Drton, M., and Uhler, C. Unpaired multi-domain causal representation learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Suo, C., Dann, E., Goh, I., Jardine, L., Kleshchevnikov, V., Park, J.-E., Botting, R. A., Stephenson, E., Engelbert, J., Tuong, Z. K., et al. Mapping the developing human immune system across organs. *Science*, 376(6597):eabo0510, 2022.

Tong, Q. and Kobayashi, K. Entropy-regularized optimal transport on multivariate normal and q-normal distributions. *Entropy*, 23(3):302, March 2021. ISSN 1099-4300.

Tran, H. T. N., Ang, K. S., Chevrier, M., Zhang, X., Lee, N. Y. S., Goh, M., and Chen, J. A benchmark of batch-effect correction methods for single-cell rna sequencing data. *Genome biology*, 21:1–32, 2020.

Tschannen, M., Bachem, O., and Lucic, M. Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069*, 2018.

Vallender, S. Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786, 1974.

Vinh, N., Epps, J., and Bailey, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11:2837–2854, 10 2010.

Virshup, I., Rybakov, S., Theis, F. J., Angerer, P., and Wolf, F. A. anndata: Annotated data. *BioRxiv*, pp. 2021–12, 2021.

Voltan, R., Secchiero, P., Corallini, F., and Zauli, G. Selective induction of tp53i3/p53-inducible gene 3 (pig3) in myeloid leukemic cells, but not in normal cells, by nutlin-3. *Molecular Carcinogenesis*, 53(6):498–504, 2014.

Wang, J., Agarwal, D., Huang, M., Hu, G., Zhou, Z., Ye, C., and Zhang, N. R. Data denoising with transfer learning in single-cell transcriptomics. *Nature Methods*, 16(9):875–878, August 2019. ISSN 1548-7105.

Weinberger, E., Lin, C., and Lee, S.-I. Isolating salient variations of interest in single-cell data with contrastivevi. *Nature Methods*, 20(9):1336–1345, 2023.

Wu, Y., Price, L. C., Wang, Z., Ioannidis, V. N., Barton, R. A., and Karypis, G. Variational causal inference. *arXiv preprint arXiv:2209.05935*, 2022.

Xu, C., Lopez, R., Mehlman, E., Regier, J., Jordan, M. I., and Yosef, N. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular systems biology*, 17(1):e9620, 2021.

Yin, K.-J., Hamblin, M., Fan, Y., Zhang, J., and Chen, Y. E. Krüpple-like factors in the central nervous system: novel mediators in stroke. *Metabolic brain disease*, 30:401–410, 2015.

Yin, M., Wang, Y., and Blei, D. M. Optimization-based causal estimation from heterogenous environments. *arXiv preprint arXiv:2109.11990*, 2021.

Yong, L., Zhou, F., Tan, L., Ma, L., Liu, J., HE, Y., Yuan, Y., Liu, Y., Zhang, J. Y., Yang, Y., and Wang, H. Continuous invariance learning. In *The Twelfth International Conference on Learning Representations*, 2024.

YosefLab. GitHub - YosefLab/scib-metrics: Accelerated, Python-only, single-cell integration benchmarking metrics — github.com. https://github.com/yoseflab/scib-metrics, 2024. [Accessed 22-05-2024].

Zhang, S., Xie, L., Cui, Y., Carone, B. R., and Chen, Y. Detecting fear-memory-related genes from neuronal scrna-seq data by diverse distributions and bhattacharyya distance. *Biomolecules*, 12(8):1130, August 2022. ISSN 2218-273X.

Zhang, Y. and Davison, B. D. Adversarial continuous learning in unsupervised domain adaptation. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part II*, pp. 672–687. Springer, 2021.

Zhang, Z., Zhao, X., Bindra, M., Qiu, P., and Zhang, X. scdisinfact: disentangled learning for integration and prediction of multi-batch multi-condition single-cell rna-sequencing data. *Nature Communications*, 15(1):912, 2024.

A. Related Work

Models like single-cell Variational Inference (scVI) facilitate data integration by incorporating environmental variables such as experimental batches or sequencing protocols alongside gene data, using one-hot vectors processed through a conditional variational autoencoder (cVAE) to reduce technical noise (Lopez et al., 2018). Its extension, single-cell ANnotation using Variational Inference (scANVI), builds on this by introducing cell annotations in a semi-supervised approach, thus enhancing cell integration across diverse environments and adeptly capturing cell type variations (Xu et al., 2021; De Donno et al., 2023). Despite their effective integration, these models may over-correct, adjusting biological signals while targeting technical noise, which can obscure subtle biological variations such as inter-patient differences or treatment effects. Moreover, these methodologies tend to aggregate all sources of spurious correlations indiscriminately, failing to discern the unique characteristics of each source (Lopez et al., 2018; De Donno et al., 2023; Tran et al., 2020). This approach inadequately addresses the nuanced interactions between these sources and biological signals, particularly problematic with continuous spurious covariates such as age or drug dosage. Models equipped to continuously adapt to these subtle variations are thus essential, ensuring that biological insights derived from single-cell genomics are not confounded by these varying conditions.

Several models in single-cell genomics have explored creating multiple latent spaces to handle different sources of variability distinctly. For instance, contrastiveVI models each covariate separately, developing a shared latent space for the common variability across covariates and an exclusive latent space for the target covariate's unique variability (Weinberger et al., 2023). Similarly, single cell disentangled Integration preserving condition-specific Factors (scDisInFact) develops a shared latent space specifically designed to account for and eliminate batch effects, while simultaneously maintaining separate latent spaces for other covariates, isolating and preserving the variations from batch influences. (Zhang et al., 2024). Yet, none of these approaches offer a control latent space dedicated to retaining batch effects while filtering out the influences of other covariates, essential for accurately distinguishing between variations caused by batch effects and those arising from true biological differences. Such methods draw inspiration from broader approaches focused on fair and disentangled representation, such as Flexibly Fair VAE (FFVAE) and Fader networks, and unsupervised disentanglement techniques such as Total Correlation VAE (β-TCVAE) (Shamsaie et al., 2024; Oh et al., 2022; Lample et al., 2018; Chen et al., 2018). The cell optimal transport model (CellOT) uses optimal transport (OT) methods to align cells from control and perturbed conditions, but its non-generative, single-covariate focus limits broader applicability (Bunne et al., 2023). Biolord offers a unique approach to supervised disentanglement, yet it faces scalability issues due to per-cell optimization (Piran et al., 2024). The invariant VAE (inVAE) method introduces conditional priors within the VAE framework to effectively disentangle spurious and invariant correlations. While it offers nuanced disentanglement, inVAE faces optimization challenges, particularly in large datasets, and does not separate latent representations for individual covariates, and does not support continuous covariates naively limiting its ability to analyze complex interactions between various biological conditions in detail (Aliee et al., 2023). On the other hand, Compositional Perturbation Autoencoder (CPA) handles drug perturbations and produce latent embedding but their assumption of linearity in the latent space limits capturing complex, non-linear biological interactions (Lotfollahi et al., 2023; Inecik et al., 2022).

While existing approaches in single-cell genomics have notably advanced the disentanglement of spurious and invariant correlations, they predominantly excel within narrowly defined scenarios. Many models, however, simplify continuous covariates by categorizing them, which undermines the granularity of biological insights and limits their applicability in precision medicine. Beyond this, there's a critical need for models that not only handle the diversity of single-cell data but also scale efficiently and train effectively given the heterogeneity inherent in these datasets. Despite the innovative nature of these methods, they are often tailored to specific experimental conditions rather than offering a universal solution across the diverse landscape of single-cell analysis. There remains an unmet need for a comprehensive model that excels in data integration, out-of-distribution prediction, and serves as a robust platform for addressing intricate biological questions across various conditions and experimental setups.

B. Supplementary Figures

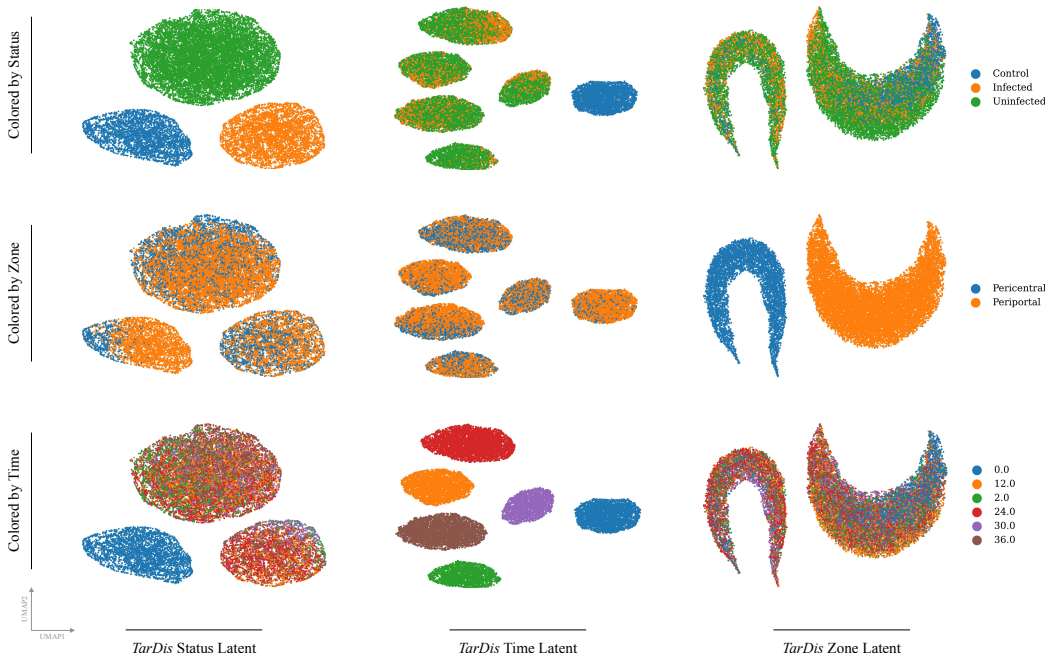


Figure 6. UMAP visualization of $TarDis_{\text{multiple}}$ latent space representations from the *Afriat* dataset. The *TarDis* model training produces four distinct latent spaces: unreserved, status, zone, and time. The UMAP plots for the status, zone, and time latent subspaces illustrate a well-structured separation of the covariates, indicating effective encoding of the underlying data distributions and disentangled relationships within these subspaces.

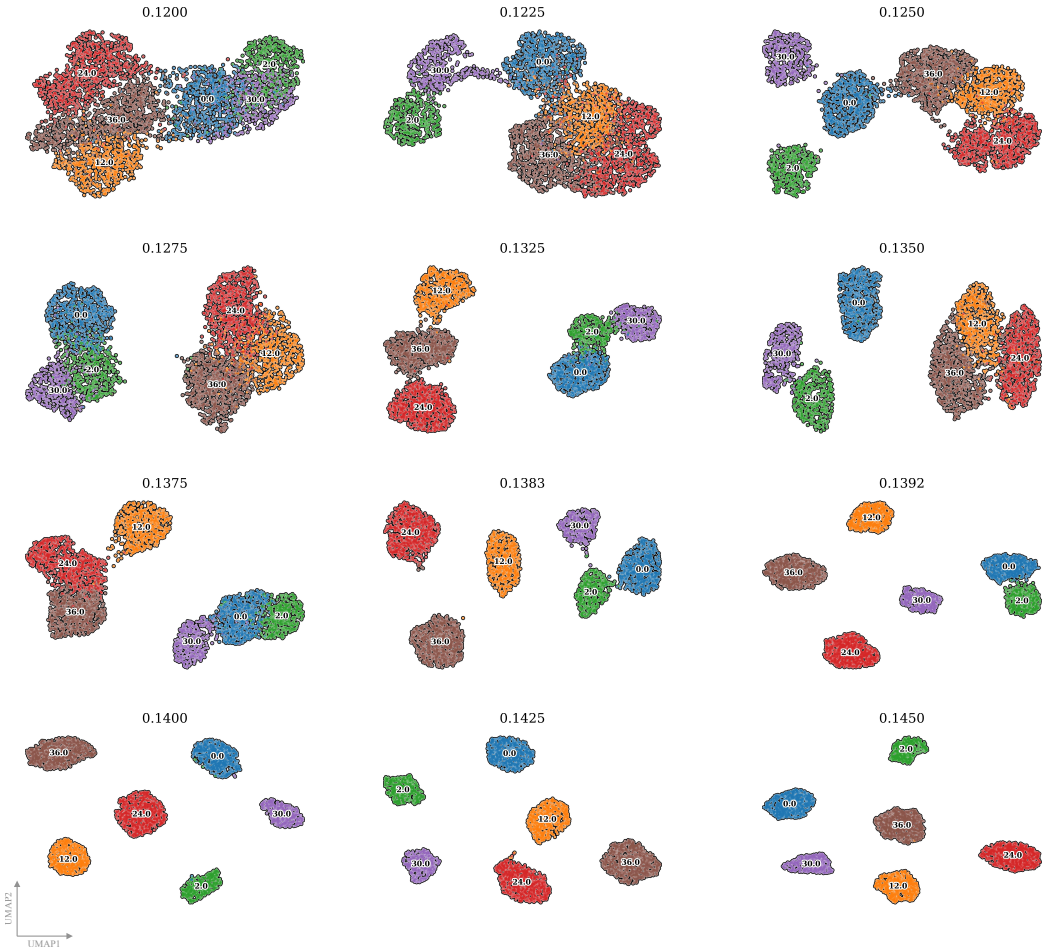


Figure 7. UMAP visualizations of disentangled time representation of *Afriat* dataset in the *TarDis* model with varying weights of the auxiliary loss λ_C. Each panel illustrates the latent space representation of targeted time covariate, highlighting how different λ_C values influence the clustering and separation of data points corresponding to different time points. As λ_C increases, given above of the UMAP visualizations, the disentanglement quality improves, evidenced by more distinct clusters, indicating the model's enhanced ability to preserve temporal information while disentangling other covariates. These visualizations provide qualitative support for the quantitative findings on the impact of auxiliary loss weight on disentanglement performance.

C. Datasets

Please be aware that this section contains embedded hyperlinks, which are essential for accessing the referenced datasets and additional resources. For optimal functionality and ease of navigation, it is highly recommended to consult the PDF version of this document. The PDF format ensures that all hyperlinks are active and can be directly accessed, facilitating seamless retrieval of the associated data and supplementary information.

C.1. *Afriat* Dataset

Description: The *Afriat* dataset, named after the first author of the study, provides high-resolution single-cell RNA sequencing and single-molecule transcript imaging data of host and parasite gene expression during the liver stage of the rodent malaria parasite *Plasmodium berghei* ANKA. It highlights spatial differences in gene expression across hepatocyte lobule zones, revealing insights into the molecular interactions between host and parasite (Afriat et al., 2022).

Number of Samples: The dataset comprises 19,053 individual cells.

Number of Features: It encompasses expression profiles across 8,203 genes.

Source: The data is publicly accessible. The raw dataset can be found under GEO accession number GSE181725. Processed data are available as a Seurat object (Butler et al., 2018) at Zenodo. The AnnData (Virshup et al., 2021) format, utilized in this study, was downloaded from Figshare, as prepared by Biolord study (Piran et al., 2024). No preprocessing or subsetting was performed on our part.

C.2. *Suo* Dataset

Description: Named after a co-author of the originating study, the *Suo* dataset offers a multi-organ, single-cell transcriptomic perspective, capturing dynamic immune system developments across nine prenatal human tissues during embryonic stages. This comprehensive dataset details the temporal and spatial maturation of immune cells, highlighting embryonic developmental timing and the interaction between different organ systems in shaping the immune landscape (Suo et al., 2022).

Number of Samples: From an initial count of 908,178 individual cells, 841,922 cells met quality control standards set by established single-cell best practices (Heumos et al., 2023).

Number of Features: The dataset, which initially profiled 33,538 genes, has been refined to focus on 8,192 highly variable genes (HVGs), following established single-cell sequencing best practices (Heumos et al., 2023).

Source: Processed data are available in AnnData format, accessible at Cellatlas portal. Additional metadata with more detailed annotation is available through the cellxgene server (Biology et al., 2023). The metadata was then refined and corrected for errors by the authors.

C.3. *Braun* Dataset

Description: Named for the first author, the Braun dataset provides a comprehensive single-cell transcriptomic analysis of the human brain during the crucial first trimester. Spanning 5 to 14 postconceptional weeks across 26 brain specimens, the dataset includes over 1.66 million cells dissected into 111 distinct biological samples. This extensive dataset captures the early spatial and transcriptional blueprint of brain development, with detailed insights into neuronal and glial differentiation trajectories (Braun et al., 2023).

Number of Samples: From an initial count of 1,665,937 individual cells, 1,661,498 cells met quality control standards set by established single-cell best practices (Heumos et al., 2023).

Number of Features: The dataset, which initially profiled 59,459 genes, has been refined to focus on 8,192 highly variable genes (HVGs), following established single-cell sequencing best practices (Heumos et al., 2023).

Source: Raw sequencing data are available from the European Genome Phenome Archive under the accession number EGAS00001004107). The data can be browsed interactively at SciLifeLab Portal and cellxgene server. The metadata was then refined and corrected for errors by the authors.

C.4. *Miller* Dataset

Description: The Miller dataset, named after the first author of the paper, provides a detailed single-cell mRNA sequencing atlas of human lung development from 11.5 to 21 weeks, integrated with studies on homogeneous human bud tip organoid cultures. This dataset specifically investigates the role of SMAD signaling in the differentiation of bud tip progenitors into airway lineages, showcasing how in vitro conditions mirror in vivo airway structures and function. This comprehensive atlas underscores critical insights into the cellular mechanisms guiding human airway differentiation (Miller et al., 2020).

Number of Samples: From an initial count of 8443 individual cells, 7405 cells met quality control standards set by established single-cell best practices (Heumos et al., 2023).

Number of Features: The dataset, which initially profiled 36,601 genes, has been refined to focus on 8,192 highly variable genes (HVGs), following established single-cell sequencing best practices (Heumos et al., 2023).

Source: The raw scRNA-seq data associated with this study are available in the EMBL-EBI ArrayExpress database under accession number E-MTAB-8221. The metadata was then refined and corrected for errors by the authors.

C.5. *Sciplex* Dataset

Description: The *Sciplex* dataset, derived from the sci-Plex technology using nuclear hashing, quantifies transcriptional responses to chemical perturbations at single-cell resolution. Applied to three cancer cell lines and exposing them to 188 distinct compounds, it evaluates dose-dependent effects and different drug responses. This high-throughput chemical screen profiles approximately 650,000 single-cell transcriptomes across about 5000 samples in a single experiment, revealing cellular heterogeneity in drug response, commonalities within compound families, and nuanced differences within compound types, particularly histone deacetylase inhibitors (Srivatsan et al., 2020).

Number of Samples: The dataset comprises 14,811 individual cells.

Number of Features: It encompasses expression profiles across 4999 genes.

Source: Both processed and raw data are accessible via NCBI GEO under accession number GSE139944. The dataset used, in its preprocessed and subsetted format, aligns with the methodology described in the CPA paper (Lotfollahi et al., 2023), provided courtesy of the authors of CPA. No further preprocessing or subsetting was conducted by our team.

C.6. *Norman* Dataset

Description: Named for the first author, the *Norman* dataset leverages high-content Perturb-seq (single-cell RNA-sequencing pooled CRISPR screens) to explore cellular and organismal complexity through combinatorial gene expression. The dataset features transcriptional responses from 284 different single or double gene knockouts, allowing for the exploration of genetic interactions at scale. This includes the mapping of regulatory pathways, classification of genetic interactions such as suppressors, and the mechanistic study of synergistic effects, notably between `CBL` and `CNN1` in erythroid differentiation (Norman et al., 2019).

Number of Samples: The dataset comprises 108,497 individual cells.

Number of Features: It encompasses expression profiles across 5000 genes.

Source: Raw data is accessible via NCBI GEO under accession number GSE133344. The dataset used, in its preprocessed and subsetted format, aligns with the methodology described in the CPA paper (Lotfollahi et al., 2023), provided courtesy of the authors of CPA. No further preprocessing or subsetting was conducted by our team.

D. Limitations

While *TarDis* introduces significant advancements in disentangling complex covariate structures in single-cell genomics, it is important to acknowledge several inherent limitations. *TarDis* operates under a supervised learning paradigm, which necessitates access to pre-labeled covariates. This requirement limits its applicability to datasets where such labels are readily available and accurately annotated, constraining its utility in less structured environments.

A notable limitation of *TarDis* is the potential for overfitting. Although rigorous validation protocols and robust regularization strategies, including elevated dropout rates and weight decay—more aggressive than those utilized in generic VAE models like scVI—are employed, the risk remains. In our study, the hyperparameters were carefully optimized at the onset of all experiments, ensuring consistent conditions across all tests, which mitigated the concerns of overfitting. It is important to note that our successful one-time optimization and the avoidance of overfitting in single-cell genomics data do not guarantee similar outcomes across other data types, hence users must conduct cautious benchmarking on validation splits to ensure the model's generalizability.

Moreover, the disentanglement of interdependent covariates introduces unique challenges. For example, accurately disentangling *age* and *donor* in a single-cell genomics data as covariates requires the presence of multiple donors of varying ages to prevent the model from conflating these factors. Without such diversity, the model risks inaccurately attributing the influence of one covariate to another, thereby undermining the reliability of the disentanglement, particularly evident in our validation splits.

Additionally, the implementation of *TarDis* introduces computational overhead, slightly slowing down the processing speed. Nevertheless, this does not significantly impact performance, even with large datasets like the *Braun* dataset, which comprises 1.6 million cells. The primary bottleneck arises from the selection of counteractive minibatches for each covariate during training, which is quantified to increase the average training time by approximately 1.8 times in comparison to scVI, when three covariates were targeted.

The encoding of covariates in a one-hot format, $\mathbf{s}_n$, while optional as mentioned in Section 2, generally fosters better disentanglement in the validation splits. However, the dependency of disentanglement on the input space may necessitate further optimization. This adjustment is crucial for enhancing the model's utility in specific downstream tasks, as demonstrated in our analysis using the *Norman* dataset in Section 3.5.

Lastly, *TarDis* necessitates numerous hyperparameters, especially concerning the loss weights for each of the four terms associated with every covariate. This complexity was manageable in our experiments through our aforementioned one-time optimization, and it did not present issues for single-cell data. However, adapting the model to new datasets could necessitate further tuning, potentially complicating its application across varied contexts. It is also important to underscore the model assumptions in Appendix E, as these foundational assumptions highlight potential limitations and areas where *TarDis* might encounter challenges.

E. Theoretical Assumptions

- Gene Dependency: The model implicitly assumes that the expression of genes can be considered independently (conditional on the latent space and covariates) when calculating losses. However, genes often exhibit co-expression or are co-regulated, which the model might not account for without specific modifications.
- Homogeneity of Cell Populations: It's implicitly assumed that cell populations are homogeneous within groups defined by covariates, which might not be the case in heterogeneous biological conditions such as tumors or developing tissues.
- Distribution of Gene Expression Counts: The model assumes that gene expression counts can be modeled effectively using a Negative Binomial distribution. This assumption is common but might not always capture the real variability and distribution in different types of datasets.
- Linearity and Gaussianity of Latent Space: The auxiliary loss assumes a Gaussian distribution for the latent vectors $\mathbf{z}_{nk}$. This implies assumptions about linearity and normality in the latent space, which may not hold in more complex or non-linear biological data structures. This assumption is critical for the model's simplicity and tractability:

$$\mathbf{z}_{nk} \sim \mathcal{N}(\boldsymbol{\mu}_{nk}, \boldsymbol{\Sigma}_{nk}) \tag{9}$$

- Static Covariate Definition: The model assumes static and well-defined positive or negative sample definitions in terms of covariate values. This is critical for the stability of the training process: $\mathbf{s}_{nk}^{(k)+}$ and $\mathbf{s}_{nk}^{(k)-}$ are fixed and consistent throughout the dataset.
- Consistency and Availability of Covariate Labels: Consistent and accurate labeling of covariates across all cells is required. Incomplete or inaccurate labels can undermine the model's effectiveness:

$$p(\mathbf{s}_{nk} = \mathbf{s}') = 1 \qquad \forall n \in N_C \tag{10}$$

- Smoothness of Latent Space: The auxiliary loss assumes the latent space is smooth and continuous, allowing for meaningful interpolation and extrapolation:

$$\forall \mathbf{z}_{nk},\ \exists \text{ continuous function } g \text{ such that } g(\mathbf{z}_{nk}) = \mathbf{x}_n \tag{11}$$

- Sensitivity to Outliers: The model does not explicitly account for outliers, which can skew learned representations. It's assumed that:

$$p(\mathbf{x}_n \text{ is outlier}) = 0 \tag{12}$$

- Assumption of Sufficient Sample Size: The effectiveness of the model in disentangling and accurately representing biological phenomena is contingent upon having a sufficiently large number of samples to cover the variability and complexity of the data. Small sample sizes could lead to overfitting and poor generalization to new data:

$$\min_k \left(\sum_{n \in N_C} \mathbb{I}(\mathbf{s}_{nk} = \mathbf{s}') \right) \geq \text{threshold} \tag{13}$$

- Data Sparsity: The model assumes it can handle sparsity in single-cell genomic data without additional modifications.
- Consistency of Environmental and Experimental Conditions: It's assumed that all cells are subject to similar environmental and experimental conditions, aside from the controlled variations represented by covariates. Variability in these conditions could introduce unmodeled noise and bias.

F. Loss Functions

Without loss of generality, various choices for the loss function are investigated, focusing on elucidating the loss incurred between the anchor point $\mathbf{x}_{nk}$ and the positive sample $(\mathbf{x}_{nk}^{(k)})^+$. The loss between the anchor point and the negative sample $(\mathbf{x}_{nk}^{(k)})^-$ can be derived similarly, with appropriate adjustments to maximize this loss.

F.1. Mean Squared Error (MSE)

The MSE between the latent representation of the anchor $\mathbf{z}_{nk}$ and its positive counterpart $(\mathbf{z}_{nk}^{(k)})^+$ for the kth covariate is given by:

$$(\mathcal{L}_{\mathrm{C}}^{(k)})_i^+(\phi; \mathbf{x}_n, \mathbf{s}_n) = \mathrm{MSE}(\mathbf{z}_{nk}, (\mathbf{z}_{nk}^{(k)})^+) \quad = \frac{1}{|\mathbf{z}_{nk}|} \sum_{j=1}^{|\mathbf{z}_{nk}|} (z_{nkj} - (z_{nkj}^{(k)})^+)^2 \tag{14}$$

However, minimizing the L_2 distance between normal vectors from distinct multivariate normal distributions with unique diagonal covariance matrices does not inherently ensure the convergence of their distributions. While this minimization may align distribution means, it disregards differences in variances and higher-order moments essential for comprehensive distributional characterization.

Mathematically speaking, if $\mathbf{z}_{nk} \sim \mathcal{N}(\boldsymbol{\mu}_{nk}, \boldsymbol{\Sigma}_{nk})$ and $(\mathbf{z}_{nk}^{(k)})^+ \sim \mathcal{N}((\boldsymbol{\mu}_{nk}^{(k)})^+, (\boldsymbol{\Sigma}_{nk}^{(k)})^+)$, by using linearity of expectation and properties of the transpose, the expected squared L_2 distance between $\mathbf{z}_{nk}$ and $(\mathbf{z}_{nk}^{(k)})^+$ can be simplified to:

$$\mathbb{E}\left[\|\mathbf{z}_{nk} - (\mathbf{z}_{nk}^{(k)})^+\|_2^2\right] = \mathbb{E}\left[\mathbf{z}_{nk}^T \mathbf{z}_{nk}\right] - \mathbb{E}\left[(\mathbf{z}_{nk})^T (\mathbf{z}_{nk}^{(k)})^+\right] - \mathbb{E}\left[((\mathbf{z}_{nk}^{(k)})^+)^T \mathbf{z}_{nk}\right] + \mathbb{E}\left[((\mathbf{z}_{nk}^{(k)})^+)^T (\mathbf{z}_{nk}^{(k)})^+\right] \tag{15}$$

For any vector $\mathbf{z}_{nk}$ with mean $\boldsymbol{\mu}_{nk}$ and covariance $\boldsymbol{\Sigma}_{nk}$, the following identity holds:

$$\mathbb{E}\left[\mathbf{z}_{nk}^T \mathbf{z}_{nk}\right] = \mathrm{tr}(\boldsymbol{\Sigma}_{nk}) + \boldsymbol{\mu}_{nk}^T \boldsymbol{\mu}_{nk} \tag{16}$$

Applying this to $(\mathbf{z}_{nk}^{(k)})^+$ and also knowing $\mathbf{z}_{nk}$ and $(\mathbf{z}_{nk}^{(k)})^+$ are independent, we have:

$$\mathbb{E}\left[((\mathbf{z}_{nk}^{(k)})^+)^T (\mathbf{z}_{nk}^{(k)})^+\right] = \mathrm{tr}((\boldsymbol{\Sigma}_{nk}^{(k)})^+) + ((\boldsymbol{\mu}_{nk}^{(k)})^+)^T (\boldsymbol{\mu}_{nk}^{(k)})^+ \tag{17}$$

$$\mathbb{E}\left[\mathbf{z}_{nk}^T (\mathbf{z}_{nk}^{(k)})^+\right] = \boldsymbol{\mu}_{nk}^T (\boldsymbol{\mu}_{nk}^{(k)})^+ \tag{18}$$

$$\mathbb{E}\left[((\mathbf{z}_{nk}^{(k)})^+)^T \mathbf{z}_{nk}\right] = ((\boldsymbol{\mu}_{nk}^{(k)})^+)^T \boldsymbol{\mu}_{nk} \tag{19}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. Substituting back, we find:

$$\mathbb{E}\left[\|\mathbf{z}_{nk} - (\mathbf{z}_{nk}^{(k)})^+\|_2^2\right] = \mathrm{tr}(\boldsymbol{\Sigma}_{nk}) + \boldsymbol{\mu}_{nk}^T \boldsymbol{\mu}_{nk} - 2\boldsymbol{\mu}_{nk}^T (\boldsymbol{\mu}_{nk}^{(k)})^+ + \mathrm{tr}((\boldsymbol{\Sigma}_{nk}^{(k)})^+) + ((\boldsymbol{\mu}_{nk}^{(k)})^+)^T (\boldsymbol{\mu}_{nk}^{(k)})^+ \tag{20}$$

To simplify further, recognizing the vector identity $\|\Delta\|_2^2 = \Delta^T \Delta$ for squared terms where $\Delta = (\boldsymbol{\mu}_{nk}^{(k)} - (\boldsymbol{\mu}_{nk}^{(k)})^+)$:

$$\mathbb{E}\left[\|\mathbf{z}_{nk} - (\mathbf{z}_{nk}^{(k)})^+\|_2^2\right] = \mathrm{tr}(\boldsymbol{\Sigma}_{nk}) + \mathrm{tr}((\boldsymbol{\Sigma}_{nk}^{(k)})^+) + \|\Delta\|_2^2 \tag{21}$$

This expression reveals that the expected squared L_2 distance depends on both the aggregate covariances and the squared difference between the means. Minimizing this distance reduces the mean disparity term $\|\Delta\|_2^2$, but does not necessarily

minimize the covariance term $\text{tr}(\mathbf{\Sigma}_{nk} + (\mathbf{\Sigma}_{nk}^{(k)})^+)$, which reflects distributional variability. However, it is crucial to ensure the convergence of our latent representations of similar pairs across their entire characteristics. Notably, as Tong & Kobayashi (2021) demonstrated, differences in the diagonal covariances of multivariate normal distributions can significantly influence the optimal transport cost and Wasserstein distance, even when the means are aligned. This highlights the importance of considering both mean and covariance differences for accurate distribution comparison. Consequently, we redirect our focus towards statistical metrics like KL divergence, which encompass the entire distribution and provide a more comprehensive assessment of distributional convergence.

F.2. KL Divergence

Unlike the L_2 distance, which primarily measures central tendency, the KL divergence accounts for both dispersion and correlation structure. Specifically, KL divergence is sensitive to differences in the means and covariance matrices of the distributions, offering a comprehensive measure of how well one distribution approximates another, beyond merely the distance between their centers.

To frame our problem contextually, assume we have determined the representation of a positive data point in a lower-dimensional space, i.e., $(\mathbf{z}_{nk}^{(k)})^+$ is fixed. With this in mind, we aim to represent the anchor point to reflect its partial similarity in its corresponding latent representation $\mathbf{z}_{nk}$. Therefore, we utilize the encoder distribution of the positive sample, $q_\phi((\mathbf{z}_{nk}^{(k)})^+|(\mathbf{x}_{nk}^{(k)})^+, (\mathbf{s}_{nk}^{(k)})^+) = \mathcal{N}((\boldsymbol{\mu}_{nk}^{(k)})^+, (\mathbf{\Sigma}_{nk}^{(k)})^+)$ as the target for the current point's distribution, $q_\phi(\mathbf{z}_{nk}|\mathbf{x}_{nk}, \mathbf{s}_{nk}) = \mathcal{N}(\boldsymbol{\mu}_{nk}, \mathbf{\Sigma}_{nk})$ given that the gradients for the forward pass of the positive sample are not computed.

Based on the KL divergence between these two multivariate Gaussian distributions, the positive pair loss $(\mathcal{L}_{\text{C}}^{(k)})_i^+(\phi; \mathbf{x}_n, \mathbf{s}_n) = -D_{\text{KL}}(\mathcal{N}(\boldsymbol{\mu}_{nk}, \mathbf{\Sigma}_{nk}) \parallel \mathcal{N}((\boldsymbol{\mu}_{nk}^{(k)})^+, (\mathbf{\Sigma}_{nk}^{(k)})^+))$ can be calculated using a straightforward and efficient formula:

$$(\mathcal{L}_{\text{C}}^{(k)})_i^+(\phi; \mathbf{x}_n, \mathbf{s}_n) = \frac{1}{2}\left[\text{tr}(\text{inv}((\mathbf{\Sigma}_{nk}^{(k)})^+)\mathbf{\Sigma}_{nk}) + ((\boldsymbol{\mu}_{nk}^{(k)})^+ - \boldsymbol{\mu}_{nk})^T \text{inv}((\mathbf{\Sigma}_{nk}^{(k)})^+)((\boldsymbol{\mu}_{nk}^{(k)})^+ - \boldsymbol{\mu}_{nk}) - |\mathbf{z}_{nk}| + \log\frac{\left|(\mathbf{\Sigma}_{nk}^{(k)})^+\right|}{|\mathbf{\Sigma}_{nk}|}\right] \quad (22)$$

Here, inv(.) stands for the inverse of a matrix, $|.|$ represents the determinant of a matrix, $|\mathbf{z}_{nk}|$ is the dimensionality of the distributions, $\mathbf{\Sigma}_{nk} = \text{diag}((\sigma_{nk1})^2, \ldots, (\sigma_{nk|\mathbf{z}_{nk}|})^2)$ and $(\mathbf{\Sigma}_{nk}^{(k)})^+ = \text{diag}(((\sigma_{nk1}^{(k)})^+)^2, \ldots, ((\sigma_{nk|\mathbf{z}_{nk}|}^{(k)})^+)^2)$. Furthermore, the determination of the determinant for such matrices is simplified, requiring only the multiplication of their diagonal elements. Therefore, equation 22 becomes:

$$(\mathcal{L}_{\text{C}}^{(k)})_i^+(\phi; \mathbf{x}_n, \mathbf{s}_n) = \frac{1}{2}\sum_{j=1}^{|\mathbf{z}_{nk}|}\left[\frac{(\sigma_{nkj})^2}{((\sigma_{nkj}^{(k)})^+)^2} + \frac{((\mu_{nkj}^{(k)})^+ - \mu_{nkj})^2}{((\sigma_{nkj}^{(k)})^+)^2} - 1 + 2\log(\sigma_{nkj}^{(k)})^+ - 2\log\sigma_{nkj}\right] \quad (23)$$

We propose summing the KL divergence over all covariates k, analogous to the total correlation (TC) in the objective function of the Relevance Factor VAE (RF-VAE)(Kim et al., 2019). This approach is designed to promote independence among latent variables. Consequently, we apply this method to the KL loss term by calculating the KL divergence between each latent representation and the standard normal distribution individually, and then summing the results.

Additionally, instead of assigning a weight to each positive pair loss function with respect to covariate k and the KL divergence between its latent representation and the prior distribution (standard normal distribution), we introduce relevance indicators, $\mathbf{r}^{(k)}$ and $\mathbf{r}_j^{(0)}$ respectively. These indicators can be learned via a variational approach. They are parameterized and updated during the training process.

$$\begin{aligned} \mathbf{r}_j^{(0)} &= \mathbf{W}_j^{(0)} \cdot \mathbf{z}_{nj} + \mathbf{b}_j^{(0)} \qquad \forall j \in \{0\} \cup J_k \\ \mathbf{r}^{(k)} &= \mathbf{W}^{(k)} \cdot \mathbf{z}_{nk} + \mathbf{b}^{(k)} \qquad \forall k \in J_k \end{aligned} \quad (24)$$

Hence the primary objective function to maximize for becomes:

$$(\mathcal{L}_{\mathrm{C}})^{+}(\phi;\mathbf{x}_n,\mathbf{s}_n)=\frac{1}{|J_k|}\sum_{k\in J_k}\left[-\mathbf{r}^{(k)}D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{\mu}_{nk},\boldsymbol{\Sigma}_{nk})\parallel\mathcal{N}((\boldsymbol{\mu}_{nk}^{(k)})^{+},(\boldsymbol{\Sigma}_{nk}^{(k)})^{+}))\right]$$
$$+\frac{1}{|J_k|+1}\sum_{\forall j\in\{0\}\cup J_k}\left[-\mathbf{r}_j^{(0)}D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{\mu}_{nj},\boldsymbol{\Sigma}_{nj})\parallel\mathcal{N}(\mathbf{0},\boldsymbol{I})\right]$$

F.3. Bhattacharyya Loss

When comparing the Bhattacharyya Loss (D_B) to the KL divergence, several key distinctions arise. KL divergence can be less effective in handling outliers and noise compared to D_B, which provides a more robust measure in noisy environments (Silva et al., 2013). Studies have demonstrated that in high-dimensional data scenarios, D_B can outperform KL divergence in both clustering accuracy and robustness to data anomalies (Cao et al., 2017).

Incorporating D_B as a loss function offers several additional advantages. First, it has shown superior performance in distinguishing between different distributions, which is essential for effective novelty detection (Sintini & Kunze, 2020) and a key aspect of disentanglement. Disentangling different factors of variation in the data often requires a measure that can accurately differentiate between various underlying distributions. Thus, the superior performance of D_B in this regard directly supports its use in disentanglement tasks. In the domain of single-cell RNA sequencing (scRNA-seq), D_B has been successfully applied to detect fear-memory-related genes from neuronal data, demonstrating its ability to handle the high heterogeneity and dropout noise inherent in such datasets (Zhang et al., 2022). Furthermore, it has been integrated into k-means clustering, enhancing the efficiency and memory-saving capabilities for large-scale scRNA-seq data analysis (Baker et al., 2021). D_B is also robust to outliers and noise, ensuring more reliable and consistent results, which is crucial for noisy datasets (Moon et al., 2018). Disentangling factors of variation in noisy datasets requires a measure that can reliably handle outliers and noisy data points without compromising the integrity of the disentangled components. D_B's robustness makes it a suitable choice for such tasks. Additionally, its symmetry and comprehensive capture of distributional differences enhance the accuracy of various analytical models (Wang et al., 2019). For disentanglement, accurately capturing and separating the underlying factors of variation in the data is essential. D_B's mathematical properties ensure that it can provide a more precise and reliable measure of these differences, facilitating better disentanglement.

Therefore, we can write the positive pair loss utilizing D_B as follows:

$$\begin{aligned}(\mathcal{L}_{\mathrm{C}}^{(k)})_i^{+}(\phi;\mathbf{x}_n,\mathbf{s}_n)&=\mathrm{DB}(\mathbf{z}_{nk},(\mathbf{z}_{nk}^{(k)})^{+})\\&=\left[\frac{1}{8}((\boldsymbol{\mu}_{nk}^{(k)})^{+}-\boldsymbol{\mu}_{nk})^{T}\left(\frac{\boldsymbol{\Sigma}_{nk}+(\boldsymbol{\Sigma}_{nk}^{(k)})^{+}}{2}\right)^{-1}((\boldsymbol{\mu}_{nk}^{(k)})^{+}-\boldsymbol{\mu}_{nk})\right.\\&\quad\left.+\frac{1}{2}\ln\left(\frac{\left|(\boldsymbol{\Sigma}_{nk}+(\boldsymbol{\Sigma}_{nk}^{(k)})^{+})\right|}{2}\right)-\frac{1}{2}\ln\left(\sqrt{|\boldsymbol{\Sigma}_{nk}|\left|(\boldsymbol{\Sigma}_{nk}^{(k)})^{+}\right|}\right)\right]\\&=\frac{1}{4}\sum_{j=1}^{|\mathbf{z}_{nk}|}\frac{((\mu_{nkj}^{(k)})^{+}-(\mu_{nkj}))^2}{(\sigma_{nkj})^2+((\sigma_{nkj}^{(k)})^{+})^2}+\frac{1}{2}\sum_{j=1}^{|\mathbf{z}_{nk}|}\ln\left(\frac{(\sigma_{nkj})^2+((\sigma_{nkj}^{(k)})^{+})^2}{2\cdot\sigma_{nkj}(\sigma_{nkj}^{(k)})^{+}}\right)\end{aligned}\tag{25}$$

F.4. Mahalanobis Loss

Mahalanobis Loss (D_M) is a robust metric for quantifying the distance-like measure between a point and a distribution, or between two points within a distribution-defined space. Unlike KL divergence and D_B, D_M measures the deviation of a point from the mean of a distribution and can be extended to compare the central tendencies of two distributions.

The innovative use of D_M significantly enhances data interpretation and clustering accuracy. The DR-A model, combining a VAE with a generative adversarial network (GAN) leverages D_M for dimensionality reduction, achieving superior clustering and more precise low-dimensional representations of scRNA-seq data (Lin et al., 2020). This precision is crucial for accurately representing covariates in lower-dimensional spaces.

The scDREAMER framework integrates D_M within an adversarial VAE to tackle skewed cell types and nested batch effects, improving batch correction and preserving biological variability across heterogeneous datasets (Shree et al., 2023). Table 1 highlights that while our model excels in batch correction, there is room for improvement in biological conservation. Therefore, we can adopt D_M to measure the dissimilarity between the latent representation of the anchor point $\mathbf{z}_{nk}$ and the respective posterior distributions $\mathbf{q}_\phi((\mathbf{z}_{nk}^{(k)})^+ \mid (\mathbf{x}_{nk}^{(k)})^+, (\mathbf{s}_{nk}^{(k)})^+)$ as follows:

$$\begin{aligned}(\mathcal{L}_{\text{C}}^{(k)})_i^+(\mathbf{x}_n, \mathbf{s}_n) &= D_M(\mathbf{z}_{nk}, (\mathbf{z}_{nk}^{(k)})^+)^2 \\ &= (\sqrt{(\mathbf{z}_{nk} - (\boldsymbol{\mu}_{nk}^{(k)})^+)^T((\boldsymbol{\Sigma}_{nk}^{(k)})^+)^{-1}(\mathbf{z}_{nk} - (\boldsymbol{\mu}_{nk}^{(k)})^+)})^2\end{aligned} \tag{26}$$

The inverse covariance matrix computation simplifies to the reciprocal of each diagonal element, resulting in:

$$(\mathcal{L}_{\text{C}}^{(k)})_i^+(\phi; \mathbf{x}_n, \mathbf{s}_n) = \sum_{j=1}^{|\mathbf{z}_{nk}|} \frac{(z_{nkj} - (\mu_{nkj}^{(k)})^+)^2}{((\sigma_{nkj}^{(k)})^+)^2} \tag{27}$$

Minimizing D_M encourages $\mathbf{z}_n$ and $(\mathbf{z}_{nk}^{(k)})^+$ to be located within high-probability regions of the latent space, as defined by the Gaussian distribution. The latent representation of the positive example $(\mathbf{z}_{nk}^{(k)})^+$ serves as a reference, with all adjustments made relative to the current anchor point $\mathbf{z}_{nk}$.

F.5. Fisher Information

Fisher information can be used to measure the amount of information that a random variable $(\mathbf{z}_{nk}^{(k)})^+$ carries about the unknown parameters $\boldsymbol{\mu}_{nk}$ and $\boldsymbol{\Sigma}_{nk}$ of a probability distribution modeling $(\mathbf{z}_{nk}^{(k)})^+$. This measurement allows for a more precise identification of the most informative latent factors, leading to more interpretable representations. Because Fisher information is grounded in information theory, the resulting disentangled factors are often more meaningful and easier to understand, which is beneficial for tasks requiring human interpretability of covariates (Tschannen et al., 2018). Representations derived using Fisher information have been shown to improve performance in downstream tasks such as classification, clustering, and anomaly detection (Khemakhem et al., 2020), which is the ultimate goal of learning latent representations of single-cell RNA-seq data. Therefore, in the context of VAEs, Fisher information aids in analyzing information loss during the encoding process:

$$I_{\mu_{nkj}}(\boldsymbol{\mu}_{nk}, \boldsymbol{\Sigma}_{nk}) = \mathbb{E}_{q_\phi((\mathbf{z}_{nk}^{(k)})^+ | \mathbf{x}_n, \mathbf{s}_n)}\left[\left(\frac{\partial}{\partial \mu_{nkj}} \log q_\phi((\mathbf{z}_{nk}^{(k)})^+ \mid \mathbf{x}_n, \mathbf{s}_n)\right)^2\right] \tag{28}$$

$$I_{\sigma_{nkj}}(\boldsymbol{\mu}_{nk}, \boldsymbol{\Sigma}_{nk}) = \mathbb{E}_{q_\phi((\mathbf{z}_{nk}^{(k)})^+ | \mathbf{x}_n, \mathbf{s}_n)}\left[\left(\frac{\partial}{\partial \sigma_{nkj}} \log q_\phi((\mathbf{z}_{nk}^{(k)})^+ \mid \mathbf{x}_n, \mathbf{s}_n)\right)^2\right] \tag{29}$$

$$(\mathcal{L}_{\text{C}}^{(k)})_i^+(\phi; \mathbf{x}_n, \mathbf{s}_n) = \sum_{j=1}^{|\mathbf{z}_{nk}|}\left[I_{\mu_{nkj}}(\boldsymbol{\mu}_{nk}, \boldsymbol{\Sigma}_{nk}) + I_{\sigma_{nkj}}(\boldsymbol{\mu}_{nk}, \boldsymbol{\Sigma}_{nk})\right] \tag{30}$$

In our case, the log-likelihood function for a single observation $\mathbf{x}_n$ is given by:

$$\log\Big(q_\phi((\mathbf{z}_{nk}^{(k)})^+ \mid \mathbf{x}_n, \mathbf{s}_n)\Big) = -\tfrac{1}{2}\left[|\mathbf{z}_{nk}|\log(2\pi) + \textstyle\sum_{j=1}^{|\mathbf{z}_{nk}|}\log\sigma_{nkj}^2 + \sum_{j=1}^{|\mathbf{z}_{nk}|}\frac{((z_{nkj}^{(k)})^+ - \mu_{nkj})^2}{\sigma_{nkj}^2}\right] \tag{31}$$

For the mean parameter μ_{nkj}:

$$\begin{aligned}I_{\mu_{nkj}}(\boldsymbol{\mu}_{nk}, \boldsymbol{\Sigma}_{nk}) &= \mathbb{E}\left[\frac{\partial}{\partial \mu_{nkj}}\frac{((z_{nkj}^{(k)})^+ - \mu_{nkj})^2}{\sigma_{nkj}^2}\right] \\ &= \frac{2}{\sigma_{nkj}^2} \cdot ((z_{nkj}^{(k)})^+ - \mu_{nkj})\end{aligned} \tag{32}$$

For the variance parameter σ^2_{nkj}:

$$\begin{aligned} I_{\sigma^2_{nkj}}(\boldsymbol{\mu}_{nk}, \boldsymbol{\Sigma}_{nk}) &= \mathbb{E}\left[\frac{\partial}{\partial \sigma^2_{nkj}} \log\Big(q_\phi((\mathbf{z}^{(k)}_{nk})^+ \mid \mathbf{x}_n, \mathbf{s}_n)\Big)\right] \\ &= 2\sigma_{nkj} - 4 \cdot \frac{((z^{(k)}_{nkj})^+ - \mu_{nkj})^2}{\sigma^2_{nkj}} \end{aligned} \tag{33}$$

G. Experimental Details

G.1. Model

Table 2: Hyperparameters for model configuration: This table lists the hyperparameters used in the model configuration, including their descriptions and assigned values.

Parameter	Description	Value
n_input	Number of input features.	
n_batch	Number of batches. If 0, no batch correction is performed.	0
n_labels	Number of labels.	0
n_hidden	Number of nodes per hidden layer. Passed into Encoder and Decoder.	512
n_latent	Dimensionality of the latent space.	24 +8 * N_K
n_layers	Number of hidden layers. Passed into Encoder and Decoder.	3
n_continuous_cov	Number of continuous covariates.	0
n_cats_per_cov	A list of integers containing the number of categories for each categorical covariate.	None
dropout_rate	Dropout rate. Passed into Encoder but not Decoder.	0.25
dispersion	Flexibility of the dispersion parameter, which can be "gene", "gene-batch", "gene-label", or "gene-cell", when gene_likelihood is either nb or zinb.	"gene"
log_variational	If True, use torch.log1p on input data before encoding for numerical stability (not normalization).	True
gene_likelihood	Distribution to use for reconstruction in the generative process. ("zinb", "nb", "poisson")	"nb"
latent_distribution	Distribution for the latent space. ("normal", "ln")	"normal"
encode_covariates	If True, covariates are concatenated to gene expression prior to passing through the encoder(s).	False
deeply_inject_covariates	If True and n_layers > 1, covariates are concatenated to the outputs of hidden layers in the encoder(s) and the decoder.	True
batch_representation	Method for encoding batch information. ("one-hot", "embedding")	"one-hot"
use_batch_norm	Specifies where to use torch.nn.BatchNorm1d in the model. ("encoder", "decoder", "none", "both")	None
use_layer_norm	Specifies where to use torch.nn.LayerNorm in the model. ("encoder", "decoder", "none", "both")	"both"
use_size_factor_key	If True, use the anndata.AnnData.obs column as defined by the size_factor_key parameter in the model's setup_anndata method as the scaling factor in the mean of the conditional distribution.	False
use_observed_lib_size	If True, use the observed library size for RNA as the scaling factor in the mean of the conditional distribution.	True
library_log_means	Vector of shape (1, n_batch) of means of the log library sizes that parameterize the prior on library size.	None
library_log_vars	Vector of shape (1, n_batch) of variances of the log library sizes that parameterize the prior on library size.	None
var_activation	Callable used to ensure positivity of the variance of the variational distribution. Passed into Encoder. The default is the exponential function.	None

Parameter	Description	Value
deeply_inject_disentengled_latents	If True, deeply inject disentangled latents.	True
include_auxillary_loss	If True, include auxiliary loss.	True
beta_kl_weight	Weight for the KL divergence term in the loss function.	0.5

G.2. Training

Table 3: Hyperparameters used for optimization: It provides a comprehensive overview of the configurations necessary to monitor and enhance model performance throughout the training

Parameter	Description	Value
max_epochs	Maximum number of training epochs.	600
train_size	Proportion of data used for training.	0.8
batch_size	Number of samples per batch.	128
check_val_every_n_epoch	Frequency of validation checks in epochs.	10
limit_train_batches	Fraction of training batches to use.	1.0
limit_val_batches	Fraction of validation batches to use.	1.0
learning_rate_monitor	Monitor learning rate during training.	True
early_stopping	Enable early stopping.	False
early_stopping_patience	Number of epochs with no improvement after which training will be stopped.	150
early_stopping_monitor	Metric to monitor for early stopping.	"elbo_train"
n_epochs_kl_warmup	Number of epochs for KL divergence warmup.	600
lr	Learning rate.	1e-4
weight_decay	Weight decay (L_2 penalty).	1e-4
optimizer	Optimizer to use.	"AdamW"
reduce_lr_on_plateau	Reduce learning rate when a metric has stopped improving.	True
lr_patience	Number of epochs with no improvement after which learning rate will be reduced.	100
lr_scheduler_metric	Metric to monitor for learning rate scheduler.	"elbo_train"

G.3. Loss

Table 4. Summary of $\mathcal{L}_C$ configuration designed for covariates, namely *status control, time,* and *zone* in *TarDis*$_{\text{multiple}}$ model trained on *Afriat* dataset. It provides insights into how each covariate contributes to the overall model loss.

Configuration				**Auxiliary Losses**			
Covariate	**Res Dim**	**Target Type**	**Loss Type**	**Latent Group**	**Weight**	**Count Type**	**Opt Type**
status	8	categorical	MSE	reserved	100	$-$	max
					10	$+$	min
				completely unreserved	10	$-$	min
					100	$+$	max
time	8	categorical	MSE	reserved	100	$-$	max
					10	$+$	min
				completely unreserved	10	$-$	min
					100	$+$	max
zone	8	categorical	MSE	reserved	100	$-$	max
					10	$+$	min
				completely unreserved	10	$-$	min
					100	$+$	max

G.4. Compute Resources and System Configuration

For the computational tasks in our research, we employed *NVIDIA Tesla A100* GPUs, which feature 40 GB of high-bandwidth `HBM2` memory each. This GPU architecture is specifically designed for accelerating machine learning and high-performance computing applications, providing substantial throughput for both single and mixed-precision computations. We allocated 64 GB of GPU memory for processing large training datasets, which facilitated efficient handling of extensive computational operations without the need for frequent data swapping, thereby minimizing I/O overhead. For smaller datasets, a reduced memory allocation of 16 GB was used, which optimized resource utilization without compromising performance. On the CPU side, our computational nodes were equipped with dual *Intel Xeon Gold 6230* processors. Each processor offers 20 cores operating at a base frequency of 2.1 GHz, which can boost up to 3.9 GHz. This setup provided a robust and responsive environment for handling non-GPU-intensive tasks and managing the preprocessing and postprocessing stages of our experiments. The system's main memory configuration included 256 GB of `DDR4 RAM` per node, which was crucial for supporting the high-throughput demands of data-intensive operations, particularly when dealing with large-scale datasets and complex computational models. Computational experiments were orchestrated using an internal *SLURM* (Simple Linux Utility for Resource Management) compute cluster. We configured SLURM to efficiently allocate resources based on the demands of queued jobs, with dynamic adjustments based on priority and current load. It should be noted that the computational resources described here sufficed for all phases of the research project; the full project did not require more compute resources than those reported for the experiments.

H. Evaluation Metrics

H.1. Average Silhouette Width

The average silhouette width (ASW) (Rousseeuw, 1987b) evaluates clustering quality by measuring the relationship between within-cluster and between-cluster distances. ASW values range from -1 to 1, where -1 indicates misclassification, 0 indicates overlapping clusters, and 1 indicates well-separated clusters.

For each data point $\mathbf{x}_n$, the silhouette coefficient $\mathrm{s}(\mathbf{x}_n)$ is calculated as:

$$\mathrm{s}(\mathbf{x}_n) = \frac{d_{\text{inter}}(\mathbf{x}_n) - d_{\text{intra}}(\mathbf{x}_n)}{\max\left(d_{\text{intra}}(\mathbf{x}_n), d_{\text{inter}}(\mathbf{x}_n)\right)} \tag{34}$$

where $d_{\text{intra}}(\mathbf{x}_n)$ is the average distance from point $\mathbf{x}_n$ to all other points within the same cluster (intra-cluster distance) and $d_{\text{inter}}(\mathbf{x}_n)$ is the minimum average distance from point $\mathbf{x}_n$ to points in any other cluster (nearest-cluster distance). The overall ASW is the mean of the silhouette coefficients for all points in the dataset:

$$\text{ASW} = \frac{1}{N_C} \sum_{n=1}^{N_C} \mathrm{s}(\mathbf{x}_n) \tag{35}$$

where N_C is the total number of data points. ASW is particularly relevant in single-cell genomics for assessing how well cells cluster based on their gene expression profiles (Rousseeuw, 1987a). This metric provides an intuitive measure of clustering quality and batch mixing, crucial for understanding both biological conservation and batch effect removal. It is particularly useful in clustering-based analyses but may be sensitive to noise and outliers.

H.2. Cell Type Average Silhouette Width

Cell type average silhouette width (Cell type ASW) (Luecken et al., 2022) evaluates cell clustering quality in single-cell transcriptomics by measuring how well cells are grouped based on type labels. The silhouette coefficient for each cell is computed similarly to general ASW. To scale the ASW values between 0 and 1, the following transformation is applied:

$$\text{celltypeASW} = \frac{\text{ASW}_c + 1}{2} \tag{36}$$

where ASW_c is the ASW computed over all cell type labels c.

H.3. Batch Average Silhouette Width

Batch average silhouette width (Batch ASW) (Luecken et al., 2022) assesses the quality of batch mixing in integrated datasets, which is essential in single-cell transcriptomics to ensure that technical variations do not obscure biological signals. The silhouette coefficient for each cell, based on batch labels, is computed similarly to general ASW.

To obtain a Batch ASW score between 0 and 1, the following transformation is applied for each batch label j:

$$\text{batchASW}_j = \frac{1}{|C_j|} \sum_{\mathbf{x}_n \in C_j} \left(1 - \mathrm{s}_{\text{batch}}(\mathbf{x}_n)\right) \tag{37}$$

where C_j is the set of cells with batch label j, $|C_j|$ is the size of this set, and $\mathrm{s}_{\text{batch}}(n)$ is the silhouette coefficient for each cell n based on batch labels. The final Batch ASW score is calculated by averaging the batch ASW values across all batch labels:

$$\text{batchASW} = \frac{1}{|B|} \sum_{j \in B} \text{batchASW}_j \tag{38}$$

where B is the set of unique batch labels. A Batch ASW score closer to 0 indicates good batch mixing, meaning batch effects have been effectively corrected (Haghverdi et al., 2018).

H.4. Isolated Label F1 Score

Precision, also known as positive predictive value, gauges the proportion of correctly predicted positive instances among the total predicted positives. It's calculated by considering True Positives (TP) against False Positives (FP), following the formula:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{39}$$

In contrast, Recall, also called sensitivity or true positive rate, measures how well the model identifies actual positive instances, crucial when false negatives are costly. Its calculation focuses on TP relative to FN, given by:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{40}$$

The F1 score, a harmonic mean of precision and recall, offers a single metric balancing both aspects, with high values indicating a well-balanced model. It is calculated as:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{41}$$

Isolated Label Scores are used to evaluate the clustering and separation of cell identity labels shared by a few batches. Specifically, the isolated label F1 score, also known as the class-wise F1 score, evaluates the F1 score for individual classes and is optimized to achieve the best clustering of these isolated labels, ensuring effective integration of rare cell types. This metric is particularly valuable for handling imbalanced datasets, such as those in single-cell genomics, where it assesses the accuracy of identifying rare cell types (Sokolova & Lapalme, 2009; Luecken et al., 2022). The original scIB package typically employs a cluster-based F1 scoring method by default. However, for the sake of speed and simplicity, we are opting to use the ASW instead as implemented in `scib-metrics` package (YosefLab, 2024). The isolated label ASW measures the separation quality of these labels. These scores address the challenge of integrating rare cell types, ensuring that integration methods can effectively manage rare cell populations. However, the performance of these scores is heavily influenced by the quality of initial annotations.

H.5. Mutual Information

Mutual information (MI) quantifies the reduction in uncertainty about one variable given knowledge of another between variables in complex systems, making it a valuable measure in both theoretical analyses and practical applications (Duncan, 1970; Kraskov et al., 2004). It measures the amount of information shared between two random variables $\mathbf{z}_n^+$ and $\mathbf{z}_n^-$ as follows:

$$\text{I}(\mathbf{z}_n^+, \mathbf{z}_n^-) = p(\mathbf{z}_n^+, \mathbf{z}_n^-) \log \left(\frac{p(\mathbf{z}_n^+, \mathbf{z}_n^-)}{p(\mathbf{z}_n^+) p(\mathbf{z}_n^-)} \right) \tag{42}$$

where $p(\mathbf{z}_n^+, \mathbf{z}_n^-)$ is the joint probability distribution of $\mathbf{z}_n^+$ and $\mathbf{z}_n^-$, and $p(\mathbf{z}_n^+)$ and $p(\mathbf{z}_n^-)$ are their marginal distributions.

The value of MI is non-negative, $\text{I}(\mathbf{z}_n^+, \mathbf{z}_n^-) \geq 0$, and measures the reduction in uncertainty of $\mathbf{z}_n^+$ given $\mathbf{z}_n^-$ and vice versa. When $\text{I}(\mathbf{z}_n^+, \mathbf{z}_n^-) = 0$, the variables are statistically independent, meaning that knowing $\mathbf{z}_n^+$ does not provide any information about $\mathbf{z}_n^-$. A higher value of MI indicates a greater level of dependency between the variables.

H.6. Normalized Mutual Information

MI is influenced by dataset size and cluster entropy, complicating comparisons across datasets. Normalization techniques, which adjust MI to a standard range, typically $[0, 1]$, enable more equitable comparisons.

$$\mathrm{NMI}(\mathbf{z}_n^+, \mathbf{z}_n^-) = \frac{\mathrm{I}(\mathbf{z}_n^+, \mathbf{z}_n^-)}{\sqrt{\mathrm{H}(\mathbf{z}_n^+)\mathrm{H}(\mathbf{z}_n^-)}} \tag{43}$$

where $\mathrm{H}(\mathbf{z}_n^+)$ and $\mathrm{H}(\mathbf{z}_n^-)$ are the entropies of $\mathbf{z}_n^+$ and $\mathbf{z}_n^-$.The higher values indicate superior clustering quality (Vinh et al., 2010). In the context of single-cell genomics, the normalized mutual information (NMI) is critical for evaluating how well clusters correspond to known cell types (Luecken et al., 2022). This metric evaluates how well cell-type labels are preserved post-integration. It is often used in scenarios requiring validation of clustering results against known labels. While it provides an intuitive measure, it may not distinguish well between near-perfect and perfect clustering.

H.7. Maximum Mutual Information Gap

The maximum mutual information gap (maxMIG) is a metric designed to evaluate the disentanglement of latent variables in complex datasets where the number of covariates exceeds two, a complexity that only particular methods are equipped to manage (Shamsaie et al., 2024; Chen et al., 2018; Higgins et al., 2017; Wu et al., 2022; Kumar et al., 2017; Kim & Mnih, 2018) due to its ability to generalize and be unbiased (Chen et al., 2018; Sepliarskaia et al., 2019; Lotfollahi et al., 2023). This measure quantifies the MI between latent representations and observed covariates, focusing on how effectively these latent variables independently capture the informative characteristics of each covariate.

The maxMIG is defined for a set of latent variables $[\mathbf{z}_k]_{k=1}^{N_K}$ and corresponding covariates $[\mathbf{s}_k]_{k=1}^{N_K}$ as:

$$\mathrm{maxMIG}(\mathbf{z}_1, \ldots, \mathbf{z}_{N_K}; \mathbf{s}_1, \ldots, \mathbf{s}_{N_K}) = \frac{1}{N_K} \sum_{k=1}^{N_K} \frac{1}{\mathrm{H}(\mathbf{s}_k)} \max_{j \neq k} [\mathrm{MI}(\mathbf{z}_k, \mathbf{s}_k) - \mathrm{MI}(\mathbf{z}_k, \mathbf{s}_j)] \tag{44}$$

The maxMIG score is computed by averaging the normalized differences between the mutual information of each latent variable with its corresponding covariate and the highest mutual information with any other covariate. This focus on maximizing the information gap helps evaluate the specificity and relevance of each latent variable to its respective covariate. Higher maxMIG values suggest better disentanglement, indicating that each latent variable is more uniquely aligned with a specific covariate, thus enhancing the model's interpretability and generalizability.

H.8. Rand Index

The Rand index (RI) serves as a pivotal metric for evaluating the concordance between two clustering outcomes. It quantifies the degree of similarity by scrutinizing the allocation of data points into clusters across two distinct clustering results. Computed as the ratio of the sum of agreements to the total number of data point pairs, RI encapsulates both intra-cluster cohesion and inter-cluster separation. The formula for calculating the Rand Index is as follows:

$$\mathrm{RI} = \frac{\mathrm{TP} + \mathrm{TN}}{\binom{N}{2}} \tag{45}$$

where $N = \mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}$. While the Rand Index offers valuable insights into clustering performance, it may have limitations when dealing with varying cluster sizes or datasets with an uncertain number of clusters.

H.9. Adjusted Rand Index

The RI quantifies the proportion of agreements between the two clusterings out of all possible pairings of elements. However, because the RI does not adjust for the chance grouping of elements, the Adjusted Rand Index (ARI) (Hubert & Arabie, 1985a; Luecken et al., 2022) is often preferred, which is defined as:

$$\mathrm{ARI} = \frac{\mathrm{RI} - \text{Expected RI}}{\text{Max RI} - \text{Expected RI}} \tag{46}$$

where the Expected RI is the expected value of the RI for random clusterings and the Max RI is the maximum possible value of the RI. Mathematically, the ARI can be expressed as:

$$\text{ARI} = \frac{\sum_{ij} \binom{n_{ij}}{2} - \frac{\left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right]}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \frac{\left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right]}{\binom{n}{2}}} \tag{47}$$

where n_{ij} is the number of elements in the intersection of cluster i in X and cluster j in Y, a_i is the number of elements in cluster i of X, b_j is the number of elements in cluster j of Y, and $\binom{n}{2}$ denotes the binomial coefficient. This adjustment provides a corrected-for-chance measure, making the ARI a more reliable metric for clustering comparison.

Values of ARI above zero indicate better-than-random agreement, with a value of 1 representing perfect agreement (Hubert & Arabie, 1985b). In single-cell data analysis, ARI is useful for validating the consistency of cell type assignments across different clustering methods. This metric is key for evaluating clustering performance in the presence of noise and is commonly used to validate clustering results in datasets with known ground truth. However, it can be less intuitive to interpret compared to simpler metrics.

H.10. k-nearest neighbor Batch Effect Test

The k-nearest neighbor batch effect test (kBET) (Büttner et al., 2018; Luecken et al., 2022) assesses batch effects in high-dimensional datasets by testing the homogeneity of batch labels within the k-nearest neighbors of each data point. It evaluates whether the neighbors of a cell are more likely to come from the same batch than expected under random mixing. kBET is a robust method designed to quantify batch effects in single-cell RNA sequencing (scRNA-seq) data. To implement kBET, one first constructs a k-nearest-neighbor (kNN) graph for each cell in the dataset, using an appropriate distance metric such as Euclidean distance in a principal component analysis (PCA)-reduced space. For each cell n, the algorithm identifies its k nearest neighbors and calculates the proportion of cells from each batch within this neighborhood, denoted as p_n^j, where j indexes the batches. Under the null hypothesis of no batch effect, the expected proportion of cells from each batch should reflect the overall batch composition in the dataset, represented as q_j. The kBET then compares the observed batch proportions p_n^j with the expected proportions q_j using a statistical test, such as the Chi-square test or a permutation-based test. The test statistic for each cell n is computed as

$$\chi_n^2 = \sum_{j=1}^{|B|} \frac{(p_n^j - q_j)^2}{q_j}$$

where $|B|$ is the number of batches. The p-value associated with the Chi-square statistic indicates the likelihood that the observed batch composition within the neighborhood of cell n is consistent with the global batch composition. These p-values are aggregated across all cells to assess the overall presence of batch effects in the dataset. The kBET statistic is:

$$\text{kBET} = \frac{1}{N} \sum_{n=1}^{N} 1_{(p_n < \alpha)} \tag{48}$$

where N is the number of neighborhoods tested, p_n is the p-value from a chi-squared test, and α is the significance threshold.

This method was evaluated using peripheral blood mononuclear cells (PBMCs) from healthy donors, effectively distinguishing cell-type-specific inter-individual variability from changes in relative proportions of cell populations. kBET is crucial for evaluating the effectiveness of batch effect correction methods in single-cell transcriptomics. The kBET tool and its detailed implementation are available on the kBET GitHub repository.

H.11. Graph Connectivity

Graph connectivity evaluates whether the kNN graph of integrated data effectively connects all cells with the same identity. For each cell identity label, a subset kNN graph is created. The graph connectivity score is then computed as the average size of the largest connected component relative to the number of nodes with that cell identity (Luecken et al., 2022). This metric ensures that cells of the same type remain connected post-integration, a critical aspect for evaluating graph-based methods. Despite its importance, calculating graph connectivity can be computationally intensive for large datasets.

In single-cell genomics, graph connectivity assesses the robustness of cell interaction networks. The formula for graph connectivity is:

$$\text{Graph Connectivity} = \frac{1}{|C|}\sum_{c\in C}\frac{|LCC(G(N_c, E_c))|}{|N_c|} \tag{49}$$

where C is the set of cell identity labels, LCC$(G(N_c, E_c))$ is the largest connected component of the graph for cells with label$c$, and$|N_c|$ is the number of nodes with cell identity c.

H.12. Coefficient of determination in VAE

The R^2 Reconstruction metric, often referred to as the coefficient of determination, is a statistical measure used to evaluate the performance of VAEs in reconstructing input data. This metric quantifies how well the reconstructed outputs from a VAE approximate the original inputs, indicating the proportion of variance in the data that is captured by the model. R^2 Reconstruction is particularly useful in the evaluation of VAEs because it provides a clear metric to gauge the accuracy of data reconstructions, facilitates comparison between different VAE architectures or configurations on the same dataset, helps identify areas where the model might be lacking, guiding further refinements. This metric is critical for researchers and practitioners using VAEs to ensure that their models not only generate new data that is statistically similar to the input data but also effectively reconstruct specific instances of input data (Inecik et al., 2022; Hetzel et al., 2022).

In the context of VAEs, the R^2 Reconstruction is defined as:

$$R^2 = 1 - \frac{\sum_{n=1}^{N_C}\|\mathbf{x}_n - \hat{\mathbf{x}}_n\|^2}{\sum_{n=1}^{N_C}\|\mathbf{x}_n - \bar{\mathbf{x}}\|^2} \tag{50}$$

where $\mathbf{x}_n$ represents the original input data, $\hat{\mathbf{x}}_n$ represents the reconstructed data produced by the VAE, and $\bar{\mathbf{x}}$ is the mean of the original input data.

The R^2 value ranges from 0 to 1, where a higher value indicates that the model has effectively captured more of the variance in the input data through its reconstructions. An R^2 value of 1 signifies perfect reconstruction, whereas a value close to 0 indicates that the model performs no better than a model that would simply predict the mean of the input data for all outputs.

H.13. Coefficient of determination for Differentially Expressed Genes in VAE

In computational biology, the evaluation of VAEs reconstruction often focuses on differentially expressed genes (DEG), which show significant changes in expression under different conditions, are critical for understanding biological processes and disease mechanisms. The R^2 Reconstruction metric is adapted in this context to specifically assess how well VAEs can reconstruct the expression patterns of these DEG. Refer to Appendix H.12 for details of R^2 reconstruction score (Inecik et al., 2022; Hetzel et al., 2022).

The R^2 Reconstruction for DEG is defined as:

$$R^2_{\text{DEG}} = 1 - \frac{\sum_{n=1}^{N_C}\|\mathbf{x}_n - \hat{\mathbf{x}}_n\|^2}{\sum_{n=1}^{N_C}\|\mathbf{x}_n - \bar{\mathbf{x}}_{\text{DEG}}\|^2} \tag{51}$$

where $\mathbf{x}_n$ represents the expression levels of DEG in the original data, $\hat{\mathbf{x}}_n$ represents their reconstructed levels from the VAE, and $\bar{\mathbf{x}}_{\text{DEG}}$ is the mean expression level of DEG.

Focusing on DEG, the R^2 Reconstruction metric specifically evaluates how effectively the VAE captures the variability and regulatory patterns in gene expression that are most biologically relevant and likely to be impacted by experimental conditions. A high R^2 value indicates that the VAE has effectively learned to model the critical aspects of gene expression relevant to the study's goals.

Reconstructing differentially expressed genes is inherently more difficult yet more critical than reconstructing overall gene expression due to several factors:

(i) *Biological Relevance* DEG often carry more biological significance than stably expressed genes, directly reflecting the cellular responses to biological stimuli or disease states.

(ii) *High Variability* DEG typically exhibit high variability in expression levels, making accurate reconstruction a complex challenge that tests the model's sensitivity and precision.

(iii) *Data Reduction* By concentrating on DEG, researchers can reduce the dimensionality of the data, focusing computational resources and analytical efforts on the most informative parts of the dataset.

(iv) *Improved Sensitivity* Models tuned to capture changes in DEG can be more sensitive to subtle but biologically important changes that might be overlooked when considering all genes.

Evaluating VAE performance using the R^2 Reconstruction metric on DEG provides insights into the model's ability to handle the most critical and dynamic components of biological data, facilitating the development of more accurate and biologically informative models.

H.14. Principal Component Regression

The principal component regression (PCR) quantifies batch removal by calculating the variance contribution of the batch effect per principal component (PC) (Luecken et al., 2022). The variance contribution of the batch effect is computed as the product of the variance explained by each PC and the corresponding R^2 value from a linear regression of the batch variable onto each PC. Mathematically, it is expressed as:

$$\text{Var}(C|B) = \sum_{g=1}^{G} \text{Var}(C|\text{PC}_g) \times R^2(\text{PC}_g \mid B) \tag{52}$$

where $\text{Var}(C|\text{PC}_g)$ is the variance of the data matrix C explained by the gth principal component and $R^2(\text{PC}_g|B)$ is the coefficient of determination for the batch variable B. This metric provides a quantitative measure of batch effects, allowing for direct comparison between methods, and is essential for assessing how well integration methods remove technical variability, particularly in large-scale multi-batch studies. However, it may not fully capture non-linear batch effects.

H.15. Local Inverse Simpson's Index

The graph local inverse Simpson's index (LISI) is a metric for evaluating batch mixing (iLISI) and cell-type separation (cLISI) in integrated single-cell datasets. It uses graph-based distances and the inverse Simpson's index to measure diversity within neighborhood compositions. Scores are rescaled from 1 to the total number of batches to a range of 0 to 1, where 0 indicates minimal integration or separation, and 1 indicates optimal mixing or segregation. This metric is especially useful for graph-based integration methods and allows for cross-method comparisons, although it requires careful parameter tuning and interpretation (Korsunsky et al., 2019; Luecken et al., 2022).

cLISI assesses the integration of diverse cell types within a combined dataset. For each cell, its kNN are identified, and the composition of cell types within this neighborhood is analyzed. The diversity is quantified using the Inverse Simpson's Index:

$$D_{\text{cLISI}} = \frac{1}{\sum_{n=1}^{N_C} p_n^2} \tag{53}$$

where p_n is the proportion of the n-th cell type in the neighborhood, and N_C is the total number of distinct cell types. The average cLISI score across all cells indicates how well cell types are mixed, with high values showing effective mixing and low values indicating poor mixing.

iLISI measures dataset mixing within the local neighborhood of each cell, quantifying how well cells from different datasets are integrated. iLISI close to the number of datasets suggests good mixing, meaning datasets are well integrated where cLISI close to 1 indicates good preservation of cell types, meaning different cell types remain well separated.

Balancing iLISI and cLISI ensures datasets are integrated effectively while preserving distinct cell type identities. Graph LISI's unified measure for both batch mixing and cell-type separation makes it a valuable tool for single-cell data integration studies, providing a standardized framework for comparing integration methods and identifying optimal strategies.