Causal disentanglement for single-cell representations and controllable counterfactual generation

Author list

Yicheng Gao[1,2,*], Kejing Dong[1,2,*], Caihua Shan[4,#], Dongsheng Li[4,#], Qi Liu[1,2,3,#]

#Corresponding authors

*These authors contribute equally to this work

Corresponding authors:

Correspondence to Qi Liu[1,2,3,#] with email: qiliu@tongji.edu.cn, Dongsheng Li [4,#] with email: dongsheng.li@microsoft.com and Caihua Shan [4,#] with email: caihuashan@microsoft.com.

Affiliations

1 Tongji Hospital, Frontier Science Center for Stem Cell Research, Bioinformatics Department, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China

2 State Key Laboratory of Cardiology and Medical Innovation Center, Shanghai East Hospital, Frontier Science Center for Stem Cell Research, Bioinformatics Department, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China

3 Shanghai Research Institute for Intelligent Autonomous Systems 201804, China

4 Microsoft Research Asia, Shanghai 200232, China

Abstract

Conducting disentanglement learning on single-cell omics data offers a promising alternative to traditional black-box representation learning by separating the semantic concepts embedded in a biological process. We present CausCell, which incorporates the causal relationships among disentangled concepts within a diffusion model to perform disentanglement learning, with the aim of increasing the explainability, generalizability and controllability of single-cell data, including spatial and temporal omics data, relative to those of the existing black-box representation learning models. Two quantitative evaluation scenarios, i.e., disentanglement and reconstruction, are presented to conduct the first comprehensive single-cell disentanglement learning benchmark, which demonstrates that CausCell outperforms the state-of-the-art methods in both scenarios. Additionally, CausCell can implement controllable generation by intervening with the concepts of single-cell data when given a causal structure. It also has the potential to uncover biological insights by generating counterfactuals from small and noisy single-cell datasets.

Main

Single-cell technologies have revolutionized the field of biology by enabling analyses to be conducted at the individual cell resolution[1]. This granular perspective has revealed the vast heterogeneity within cell populations, leading to new insights into cellular functions, development processes, and diseases[2]. However, the complexity of single-cell data, which are characterized by high dimensionality and the presence of various entangled concepts, poses great challenges for data interpretation tasks. The process of disentangling these data, i.e., separating complex, intertwined signals into distinct, interpretable components, has therefore emerged as a critical

task[3] and presented to be the necessary path to build the virtual cell[4].

Disentangled representation learning seeks to capture and separate the underlying concepts embedded in intertwined data, such as images[5]. Unlike traditional end-to-end black-box representation learning methods, which often learn shortcuts by predicting human-annotated labels or reconstructing observable data, disentangled representation learning mimics the human understanding process by leveraging hidden semantic concepts to make decisions[6]. However, this is very challenging for single-cell data, as they are more complex and noisier than data in traditional machine learning communities, such as images. Additionally, these latent concepts in single-cell data are often causally connected, making it hard to clearly separate them with existing disentanglement methods. It requires more advanced techniques to capture latent concepts with causal structure and subsequently establish accurate mappings between the data and concepts in single-cell data.

Several studies have attempted to apply disentangled representation learning to obtain interpretable and manipulable representations of single-cell data. These methods can be categorized into two main groups. (1) Statistical methods: Techniques such as factor analysis or nonnegative matrix factorization are used to identify various biological programs on the basis of statistical patterns[7, 8]. However, these models do not consider the causal structures between concepts and cannot manipulate these concepts. (2) Learning-based methods: These approaches are generative and often utilize variational autoencoder (VAE)-based methods to learn hidden concepts by reconstructing single-cell data[3, 9-11]. For example, CPA decouples perturbation responses[11], scDisInFact removes batch effects[9], and Biolord disentangles the concepts contained in single-cell data[3]. However, similar to statistical methods, these methods also cannot guarantee the causality of concepts. In addition, most methods rely on latent optimization[12], which can result in a loss of fine-grained concept representations at the single-cell resolution level. Notably, prior studies have failed to design and implement comprehensive quantitative benchmarking to rigorously evaluate these disentangled methods. Collectively, a comprehensive benchmarking of existing disentanglement methods, along with the development of a causal disentanglement technique for single cell omics data, is lacking and urgently needed.

Herein, we introduce CausCell, the first deep generative framework for conducting causal disentanglement and counterfactual generation on single-cell omics data (Fig. 1a). CausCell combines a structural causal model[13, 14] (SCM) with a diffusion model, offering unprecedented advantages in three aspects for obtaining disentangled single-data representations, including for spatial and temporal omics data: (1) Explainability: CausCell leverages an SCM to recover latent concepts with semantic meanings and their causal relationships via a causal directed acyclic graph (cDAG), substantially enhancing the interpretability of the model. (2) Generalizability: Unlike previously developed VAE-based methods, CausCell uses a diffusion model as its backbone, which provides strong generative and generalization capabilities, ensuring a high-quality sample generation process[15]. (3) Controllability: CausCell enables controllable generation by manipulating disentangled representations in the latent space while preserving their consistency with the underlying causal structure. In addition, to disentangle single-cell data into various concepts, we assume that each cell is generated by two types of concepts, i.e., observed and

unexplained concepts. For example, observed concepts may involve the cancer type related to single-cell tumour omics data or spatial-domain loci derived from spatial single-cell data, and unexplained concepts are the potential unknown concepts contained in the given data. Therefore, such a framework enables us to effectively distinguish and explore the concepts hidden in the latent space. To train our model, we propose a new loss function that incorporates a new evidence lower bound (ELBO) loss and an independence constraint for the unexplained concepts. As a result, two quantitative evaluation scenarios, i.e., disentanglement and reconstruction, are presented to conduct the first comprehensive single-cell disentanglement learning benchmark, which demonstrates that CausCell outperforms the state-of-the-art tools in both scenarios. Furthermore, we validate the effectiveness of the cDAG in our model, showing that it can generate cells that are consistent with the underlying causal systems when interventions occur. Finally, we show that CausCell can uncover new biological insights when the input experimental single-cell omics dataset is small, noisy and inaccessible. Collectively, CausCell presents an unprecedented perspective and method for obtaining causal disentanglement representations of single-cell data compared with traditional black-box representation learning, and it can uncover novel and interpretable biological insights from single-cell data by controllable counterfactual generation.

A comprehensive quantitative disentanglement representation benchmark

A comprehensive quantitative disentanglement model benchmark is critical for establishing the reliability of such models, and this step was overlooked in previous studies. Evaluating the effectiveness of a disentanglement model involves the evaluation of two key aspects: (1) concept disentanglement and (2) reconstruction. The first aspect reflects the ability of the tested model to accurately capture and separate underlying semantic concepts, whereas the second aspect determines the quality and fidelity of the generated counterfactual samples by manipulating concepts for further biological analysis. Properly evaluating these aspects is essential for ensuring the practical applicability and robustness of the developed model.

To conduct comprehensive benchmarking, we collected five distinct single-cell datasets spanning different biological domains[16-20], each with various causal relationships among different concepts (Supplementary Note 1 and 2, Supplementary Fig. 1 and Supplementary Table 1). We also included two experimental settings, in-distribution (ID) and out-of-distribution (OOD) settings, on the basis of whether the different combinations of concept labels were presented during training (Fig. 2b and Supplementary Note 3). In the ID setting, the model had seen all possible combinations of concept labels during training, providing a direct assessment of its ability to operate within the same data distribution. The more challenging OOD setting involved cases where the model encountered unseen combinations of concept labels, and this task aimed to assess the ability of the model to transfer complex relationships among concept representations and examine its generalizability. Furthermore, we defined five types of metrics to evaluate various aspects of disentanglement and reconstruction performance (Supplementary Note 4). Disentanglement performance was evaluated by determining whether the learned concept representation contained sufficient information for predicting the concept labels. Therefore, we defined (1) the predictive ability of the concept embeddings and (2) the clustering consistency of the embeddings when the label granularity was varied (Supplementary Note 5). For reconstruction, we defined (3) trend matching, (4) geometric structure consistency and (5) a fine-grained score to

test the consistency of the generated single-cell data with the test data distribution.

To evaluate the disentanglement performance of the proposed approach, scDisInFact was selected as the baseline model because it uniquely maintains a single-cell resolution in concept representation tasks. Most existing disentanglement models rely on latent optimization, wherein cells sharing the same concept label are represented by identical concept representations[12]. These approaches compromise the single-cell resolution, thereby hindering the disentanglement performance assessment. Preserving concept representations at the single-cell level is crucial, as doing so captures the heterogeneity inherent within cell populations. For example, within the T-cell lineage, distinct subtypes, such as exhausted T cells and effector T cells, exhibit unique functional states. Utilizing a uniform representation for all T cells would obscure these subtle biological distinctions. Our results showed that CausCell outperformed scDisInFact in terms of various predictive metrics, including accuracy, precision, weighted F1 scores and weighted recall scores (Figs. 1c, d and Supplementary Figs. 2-5). To evaluate the concept embedding generalization capabilities of the models, we varied the granularity of their concept labels by training the models on coarse-grained cell type information and subsequently assessing their performance in terms of fine-grained cell subtype consistency (Supplementary Note 6). CausCell also demonstrated superior performance across several clustering consistency metrics, such as the normalized mutual information (NMI) and adjusted Rand index (ARI) scores (Figs. 1e, f and Supplementary Note 7).

To evaluate the reconstruction performance of the models, we used the Pearson correlation coefficient (PCC) and mean squared error (MSE) to measure the consistency of the trend between the generated data and the original data. Additionally, NMI and the ARI were employed to assess the preservation of geometric structures. A fine-grained score was also introduced to evaluate the maintenance of the marker genes in the generated cells. Then, CausCell was benchmarked against six baseline models, including four mainstream disentanglement-based models (Biolord[3], scDisInFact[9], CPA[11] and MichiGAN[10]) and two generative models without disentanglement (scVI[21] and scGen[22]) across all the above evaluation metrics. Theoretically, a trade-off exists between disentanglement and reconstruction performance[23, 24]. Our results demonstrated that CausCell not only surpassed all disentanglement-based models but also achieved reconstruction performance that was on par with or superior to that of the mainstream generative models (Fig. 1g and Supplementary Figs. 2-5). The exceptional performance of CausCell in both disentanglement and reconstruction tasks provided a solid foundation for various downstream analyses. These included generating samples that adhered to causal structures and producing reliable counterfactual samples, thereby facilitating the discovery of new biological insights.

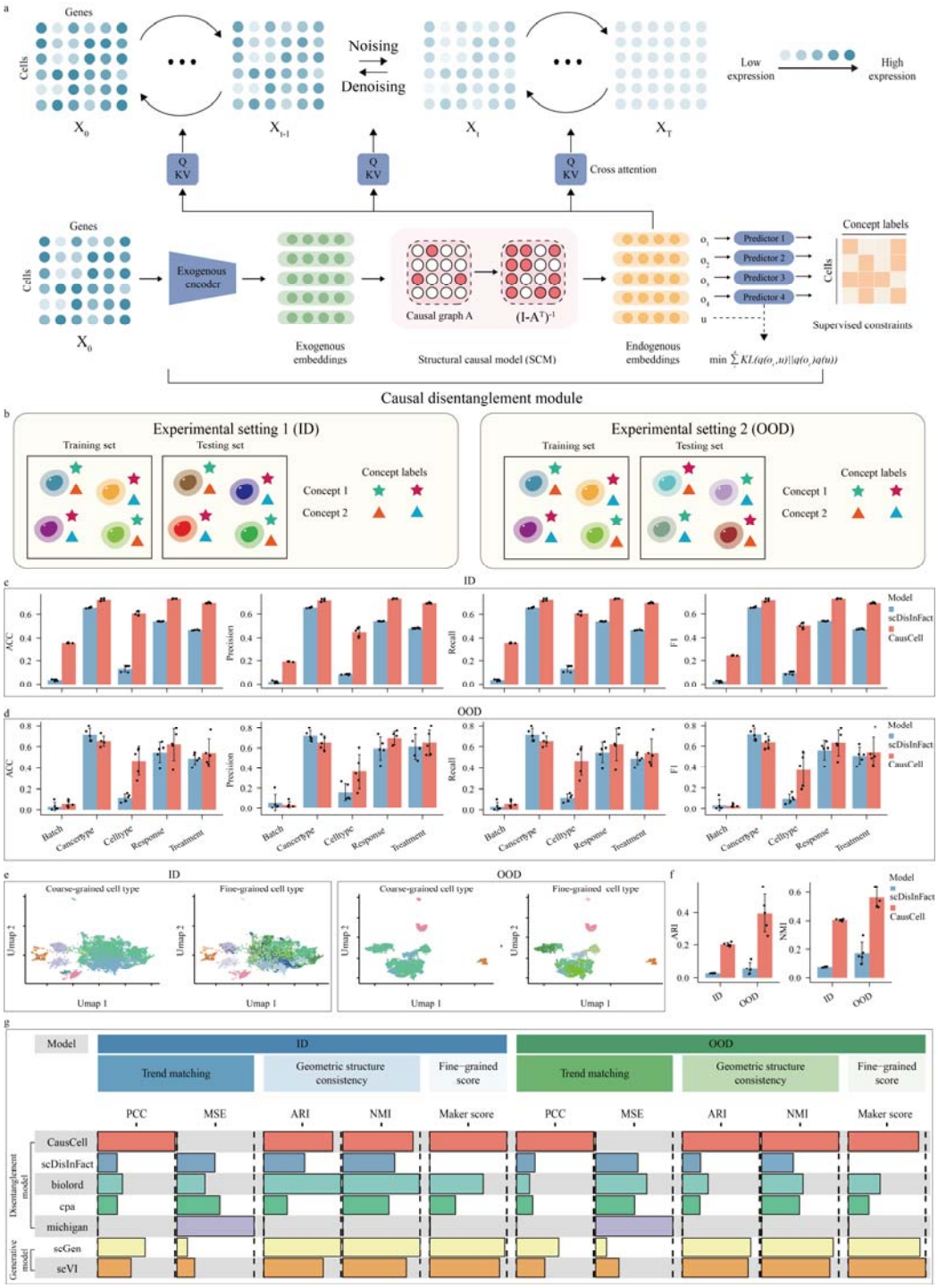

Fig. 1. Overview of CausCell and its disentanglement and reconstruction performance.

a. The framework of CausCell, which consists of a causal disentanglement module and a diffusion-based generative module. b. The two experimental settings for benchmarking the disentanglement and reconstruction capabilities of the tested model. c. The results of a disentanglement performance comparison conducted under experimental setting 1 (ID) for each concept contained in the ICI_response dataset. d. The results of a disentanglement performance comparison conducted under experimental setting 2 (OOD) for each concept

contained in the ICI_response dataset. e. UMAPs of the cell type embeddings produced with coarse-grained and fine-grained cell type annotations across the two experimental settings. f. NMI and ARI metrics attained by different models for evaluating their fine-grained geometric property preservation capabilities across the two experimental settings. g. Comparison among different models in terms of their reconstruction performance across two experimental settings. This is created by the funcky heatmap (version 0.5.0), so the lengths of the bars are proportional to the min-max normalized mean values. For all the bar charts, the data are presented as mean values, with each error bar representing the standard deviation based on n=5 (fivefold cross-validation).

Realistic biological application of CausCell

First, we demonstrated that CausCell could enhance the plausibility and consistency of the counterfactual generation process by aligning with the underlying causal structure (Fig. 2a). Counterfactual generation plays a critical role in generating hypothetical versions of single-cell data by intervening in one or several concepts, and it aims to investigate how scientific conclusions change and explore the reasons behind these changes, presenting to be an important task to build a virtual cell[4]. However, the existing disentanglement models exhibit a critical limitation because they intervene in concepts without considering their causal relationships with other concepts, resulting in the generation of unrealistic or erroneous samples. To further illustrate this point, we compared the quality levels of the samples generated by CausCell and a modified version, CausCell_IND, which lacks the causal structure and treats concept relationships as independent by removing the SCM layer from the model. Utilizing the “Control” cells contained in a spatial-temporal single cell data set, i.e., the Spatiotemporally_liver dataset, we performed interventions on the “Inject,” “Time,” and “Infect” concepts by modifying their concept labels and generating counterfactual cells with modified concept labels. We then compared the infection statuses of cells quantified by their PBAGenesScore[18] values (Supplementary Note 8). Regarding CausCell_IND, we observed that the cells generated with the “Inject” intervention presented the highest PBAGenesScore values across all time points, regardless of whether the cells were infected, even when the time value was set to 0 (Fig. 2c and Supplementary Fig. 6). This unrealistic behaviour was mitigated by incorporating the causal structure into CausCell, allowing the model to generate more realistic samples that were consistent with the underlying causal relationships, where the injected and infected cells presented higher scores than other conditions did (Fig. 2b and Supplementary Fig. 6).

Second, we validated that CausCell could reveal biological insights even from small and noisy single-cell datasets via controllable counterfactual generation (Supplementary Note 9). Single-cell data are often confounded by latent concepts, which may obscure critical biological signals, especially when the sample size is small. As a result, wet-lab experiments typically require large sample sizes and significant costs for high-throughput sequencing. The controllable counterfactual generation process conducted by CausCell offers a promising solution for augmenting data when the number of data samples is small. To illustrate this point, we analysed another spatial-temporal dataset, i.e., the mouse ageing dataset (MERFISH_brain)[17], which includes 3 mice aged 4 weeks (4wk), 3 mice aged 24 weeks (24wk), and 5 mice aged 90 weeks (90wk). A previous study[17] revealed that the expressions of 3 genes in oligodendrocytes (Oligo), 2 genes in microglia (Micro),

2 genes in astrocytes (Astro) and 1 gene in endothelial cells (Endo) increased with age, whereas the expression of one gene (*Gpc5*) in Astro decreased. However, when the sample sizes were reduced (by selecting only 2 mice from each age group), these gene expression trends were no longer observed (Fig. 2d and Supplementary Fig. 7). We subsequently used this smaller dataset to train CausCell and performed controllable counterfactual generation on the 4-week-old mice by intervening in the age concept to simulate mice at 24 and 90 weeks of age. By using these counterfactual samples, we replicated the analysis and found that 6 out of the 9 genes presented the same expression trends as previous study[17] (Fig. 2e and Supplementary Fig. 7). Additionally, the abundance of cells with high degrees of expression for 2 of the remaining 3 genes also increased (Supplementary Fig. 8).

Furthermore, we statistically analysed the number of age-related differential genes contained in different cell types across various spatial domains, which yielded results similar to those of a previous study[17] (Supplementary Fig. 9 and Supplementary Note 9). Notably, we obtained a new finding: the upregulated genes in Micro were enriched in the striatum domain, which has never been reported in previous studies[17]. A GO enrichment analysis[25] revealed that these age-related, upregulated genes were associated with multiple cell adhesion pathways, particularly those involving immune cells, in addition to immune activation pathways (Fig. 2f and Supplementary Note 10). These findings suggest that Micro may contribute to brain ageing through cell adhesion mechanisms. A further analysis identified *Lilrb4a* as the gene with the most enriched pathways, including T-cell activation and leukocyte adhesion, which were not highlighted in earlier work and could not be identified in the original data (Fig. 2h). A recent study[26] revealed that Micro in Alzheimer's disease patients express high levels of *LILRB4*, the homologous gene to *Lilrb4a*. Additionally, its family gene *LilrB3* can activate microglia into a proinflammatory state[27]. This finding highlights the potential role of *Lilrb4a* in influencing brain ageing by regulating the immune states of microglia cells. Finally, we assessed the expression of *Lilrb4a* across different cell types and spatial domains and found that its expression increased with age in both neuronal and nonneuronal cells, with spatial specificity (Figs. 2i, j). These findings suggest that *Lilrb4a* may be a common regulator of ageing across various cell types in the brains of mice.

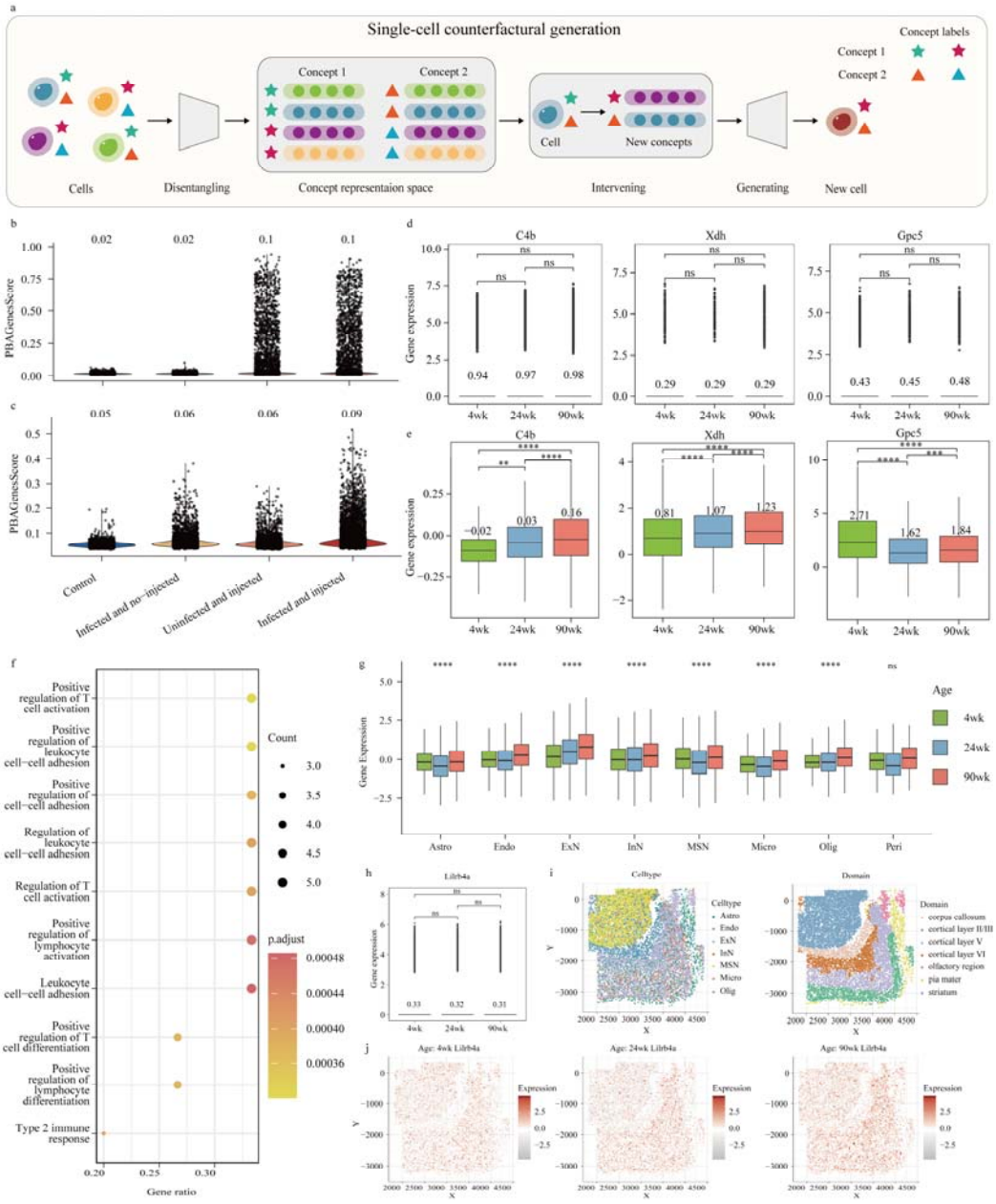

Fig. 2. The counterfactual generation results produced by CausCell.

a. Schematic diagram of the single-cell counterfactual generation process implemented by CausCell. b. PBAGenesScore values obtained for various intervention-associated cells generated by CausCell without incorporating a causal structure (CausCell_IND). c. PBAGenesScore values produced for various intervention-associated cells generated by the CausCell version incorporating the causal structure. d. Gene expression differences among different ages in the original dataset, where *C4b* in Oligo, *Xdh* in Endo and *Gpc5* in Astro. e. Gene expression differences among different ages in the counterfactually generated dataset, where *C4b* in Oligo, *Xdh* in Endo and *Gpc5* in Astro. f. GO enrichment analysis of the differentially expressed upregulated genes in microglia enriched in the striatum domain. g. Age-related *Lilrb4a* gene expression differences across all cell types. h. *Lilrb4a* expression

differences among different ages in the original dataset. i. Spatial segmentation results concerning the anatomical regions of the mouse brain. j. The spatial distribution characteristics of *Lilrb4a* expression across all ages. In the GO enrichment analysis, a hypergeometric test was performed. In each boxplot, the box boundaries represent the interquartile range, the whiskers extend to the most extreme data points within 1.5 times the interquartile range, the value indicated above the box represents the mean value, and the black line inside the box represents the median. Statistical tests were conducted via a two-sided t test with significance levels of ns: p>0.05; *: p≤0.05; **: p≤0.01; ***: p≤0.001; ****: p≤0.0001), except for subfigure G, where an analysis of variance (ANOVA) test was used.

Discussion

In conclusion, we presented comprehensive evaluation metrics that demonstrate the superior disentanglement and reconstruction performance of CausCell in single-cell disentanglement representation tasks. Additionally, by leveraging causal structures and diffusion models, CausCell has the potential to generate more realistic samples and uncover meaningful biological insights through the generation of counterfactuals, particularly when the given sample size is small. Future updating of CausCell are expected: (1) Herein, we applied the linear version of an SCM, i.e., we utilized a closed-form solution to transform concept representations given a predefined causal structure. In practice, the relationships among concepts can be nonlinearly defined in a causal structure. We need to add more learnable SCM layers to transform concepts and represent their causal relationships. (2) While our focus was on evaluating the performance of the disentanglement model and demonstrating the benefits of the causal prior in the counterfactual generation task, many potential applications of this model remain to be explored. The current CausCell framework can be naturally extended to model more modalities in single-cell data, even a foundational disentanglement model in the future.

Methods

Overview of CausCell

Consider the input data a single-cell sequencing dataset of N cells $D = \{(x^i, y^i)\}_{i=1}^N$, where x^i represents the gene expression vector for cell i and y^i consists of M observed concept labels for that cell. The disentanglement process involves learning a series of latent concept embeddings $z^i = (o_1^i, o_2^i, \ldots, o_m^i, u^i)$, where each o_j^i is a low-dimensional vector corresponding to the label y_j^i, and u^i serves as an extra captured embedding of the unobserved concepts. These embeddings z^i are then employed during the generative process to achieve a better reconstruction effect. Previously, VAEs were implemented as generative modules, but they tend to produce lower-quality samples, especially when addressing high-dimensional data such as high-resolution images data[28]. In contrast, diffusion models have been demonstrated to produce high-quality samples by generating samples in a step-by-step manner. However, their lack of an interpretable latent space makes them less suitable for disentanglement and explainability purposes. To address these issues, we propose CausCell to learn a causal disentanglement module F for concept embeddings and then integrate them into the diffusion process G. Both modules are incorporated into a unified network, which is end-to-end trained via a newly derived ELBO loss.

Causal disentanglement module *F*

The goal is to learn causal representations of concepts z^i. The disentanglement learning methods used in previous single-cell studies were constrained by the assumption of concept independence. However, concepts are not necessarily independent in practice. Instead, an underlying causal structure is present and renders these concepts dependent. Therefore, we introduce an SCM layer to construct causal relationships among the concepts. Specifically, the observed concepts o_j^i are connected through a DAG, which is represented by an adjacency matrix A. We apply a linear version of the SCM, which satisfies the following equation:

$$z = A^T z + \epsilon = (I - A^T)^{-1}\varepsilon, \varepsilon \sim N(0, I)$$

where ε represents the exogenous variables and z denotes the endogenous variables for capturing the latent concepts. We first learn a function (a multilayer perceptron (MLP)) for extracting the exogenous concept embeddings ε from the gene expression x_0 and then leverage ε to transform it into z via a closed-form solution.

To ensure the semantic meanings of the concept embeddings and to enforce causal disentanglement, we employ m discriminators $D = \{D_1, D_2, \ldots, D_m\}$, each of which corresponds to an observed concept. These discriminators are trained to predict the observed concept labels from the embeddings o_j^i via the cross-entropy loss:

$$L_F = \mathbb{E}_{x_0}[\sum_{i=1}^{m} L(D_i(o_i), y_i) - \sum_{i=1}^{m} L(D_i(u), y_i)]$$

where L is the cross-entropy loss function. This setup also encourages independence between the observed concepts o and the unexplained concept u.

Generative module *G*

We use a diffusion model as our generative backbone in CausCell because of its powerful generative capabilities. A diffusion model defines a latent variable distribution $p(x_{0:T})$ over a gene expression vector x_0 sampled from a single-cell distribution, as well as noisy gene expression vectors $x_{1:T} \coloneqq x_1, x_2, \ldots, x_T$ that represent a gradual transformation of x_0 into random Gaussian noise x_T. The reverse diffusion process is modelled as a Markov chain:

$$p(x_{0:T}) = p(x_T)\Pi_{t=0}^{T-1} p_\theta(x_t|x_{x_{t+1}})$$

where p_θ is a learned denoising distribution parameterized by a neural network with a parameter θ.

The forward diffusion process q adds Gaussian noise to x_0 at each step:

$$q(x_{1:T}|x_0) = \Pi_{t=1}^{T} q(x_t|x_{t-1})$$

with a predefined noise schedule $\{\alpha_t\}_{t=1}^{T}$. The marginal distribution can be directly computed as shown below:

$$q(x_t|x_0) = N(x_t; \sqrt{\bar{\alpha}}x_0, (1-\bar{\alpha})I)$$

where $\bar{\alpha} = \Pi_{t=1}^{T}\alpha_t$.

The denoising network g_θ in the diffusion model is an ϵ predictor in most cases. However, single-cell data often exhibit extreme sparsity, and the corrupted input at time step t is mostly pure noise. Under this setting, the model is likely to learn to reverse the noise schedule instead of the true data posterior. Therefore, we adopt the x_0-predictor[29] in this study, and a simplified training loss for the generative module is defined as follows:

$$L_G = \mathbb{E}_{x_0,z,t}\left[||x_0 - g_\theta\left(\sqrt{\overline{\alpha_t}}x_0 + \sqrt{1-\overline{\alpha_t}}\xi, z, t\right)||^2\right]$$

where ξ is the noise sampled from $N \sim (0, I)$ and t is the time step.

Integration causal disentanglement module F with a diffusion process G

The causal disentanglement module takes a single-cell expression profile as its input to obtain a set of concept embeddings z^i under the guarantee of a causal structure. We then use these concept embeddings as the condition information for the generative module. Specifically, we incorporate these concept embeddings into the reverse diffusion process of the generative model through a cross-attention mechanism. By conditioning this process on the concept embeddings, the model can generate gene expression profiles that are consistent with specific biological concepts, allowing for a controlled and interpretable data generation procedure. The cross-attention mechanism[30] facilitates this conditioning task by enabling the model to focus on the relevant parts of the concept embeddings during the generation phase. It is mathematically defined as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V$$

where Q, K, and V are the query, key, and value vectors, respectively. d is the scaling vector, which is used to ensure numerical stability in the softmax function, and its value is set as the dimensionality of the head. In this study, the corrupted gene expression profile x_t^i in the diffusion module serves as a query, and the concept embeddings z^i act as keys and values.

Evidence lower bound of CausCell

We formulate the training objective via a variational inference approach to derive the ELBO for optimizing the model parameters. We treat both z and ϵ as latent variables. Consider the following conditional generative model:

$$p(x, z, \epsilon\,|y) = p(x|z, \epsilon, y)p(\epsilon, z|y)$$

We define $p_\epsilon(\epsilon) = N(0, I)$ and the joint prior $p(\epsilon, z|y)$ for latent variables z and ϵ as follows[14]:

$$p(\epsilon, z|y) = p_\epsilon(\epsilon)p(z|y)$$

We define CausCell as a causal disentanglement-based diffusion model represented by the conditional probability distribution $p(x_{0:T}, z, \epsilon|y)$, which can be factorized as follows:

$$p(x_{0:T}, z, \epsilon|y) = p(x_T)p(z, \epsilon|y)\Pi_{t=1}^T p_\theta(x_{t-1}|x_t, z, \epsilon, y)$$

This model implements a reverse diffusion process $p_\theta(x_{t-1}|x_t, z, \epsilon, y)$ over $x_{0:T}$, which is conditioned on the endogenous variables z, the exogenous variables ϵ and the concept labels y. All of these variables are independent of the diffusion process because these variables are properties of the input, not control variables of the diffusion process.

To optimize the model parameters, we apply variational inference twice to impose a variational lower bound on the conditional log-likelihood of the concept labels $log\ p(x_o|y)$:

Proposition 1: The ELBO of CausCell can be derived as follows (Proof: Supplementary Note 11):

$$\begin{aligned} log p(x_0|y) &\geq ELBO \\ &= -KL(q_\phi(z, \epsilon|x_0, y)||p(z, \epsilon|y)) - KL(q(x_T|x_0)||p(x_T)) + \mathbb{E}_{q(x_1|x_0)}\left[\mathbb{E}_{q_\phi(z,\epsilon|x_0,y)}[p(x_0|x_1, z, \epsilon, y)]\right] \\ &\quad - \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)}\left[\mathbb{E}_{q_\phi(z,\epsilon|x_0,y)}[KL(q(x_{t-1}|x_t, x_0)|p(x_{t-1}|x_t, z, \epsilon, y)]\right] \end{aligned}$$

The above equation is intractable in general. Given an SCM with latent exogenous independent variables ϵ and the latent endogenous variables z, we have $z = A^T z + \epsilon = (I - A^T)^{-1}\epsilon$. For simplicity, we denote $C = (I - A^T)^{-1}$. Leveraging the one-to-one correspondence between ϵ

and z, we can simplify the variational posterior as follows:

$$\begin{aligned} q_\phi(\epsilon, z|x_0, y) &= q_\phi(\epsilon|x, y)\delta(z = C\epsilon) \\ &= q_\phi(z|x, y)\delta(\epsilon = C^{-1}z) \end{aligned}$$

where $\delta(\cdot)$ is the Dirac delta function. According to the model assumptions introduced above, we can further simplify the ELBO as follows:

Proposition 2: The ELBO of CausCell can be rewritten as follows (Proof: Supplementary Note 12):

$$\begin{aligned} ELBO \quad &= -KL(q_\phi(\epsilon|x_0, y)||p_\epsilon(\epsilon)) - KL(q_\phi(z|x_0, y)||p(z)) + E_{q_\phi(z|x,y)}[p(y|z)] - E_{q_\phi(z|x_0,y)}[p(y)] - KL(q(x_T|x_0)||p(x_T)) \\ &+ \mathbb{E}_{q(x_1|x_0)}\left[\mathbb{E}_{q_\phi(z|x_0,y)}[p(x_0|x_1, z)]\right] - \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)}\left[\mathbb{E}_{q_\phi(z|x_0,y)}[KL(q(x_{t-1}|x_t, x_0)||p(x_{t-1}|x_t, z)]\right] \end{aligned}$$

In practical scenarios, the latent factors z consist of observed variables o and unobserved variables u such that $z = [o, u]$. Assuming independence between o and u, we augment the ELBO with a term $-\gamma \sum_i^m KL\left(q_\phi(o_i, u)||q_\phi(o_i)q_\phi(u)\right)$ that encourages this independence, and it is implemented via an adversarial debiasing strategy[31] using discriminators. Therefore, the overall training objective is to maximize the ELBO, leading to the following combined loss function (Supplementary Note 14):

$$\mathcal{L} = L_F + L_G + KL(q_\phi(\epsilon|x_0)||p_\epsilon(\epsilon)) + KL(q_\phi(u|x_0)||p(u)) + KL(q_\phi(o|x_0, y)||p(o))$$

Benefit of disentanglement in the diffusion model

We show that the reverse diffusion process introduces an information bottleneck effect, promoting disentanglement by dynamically allocating information to the latent concepts as the time steps increases[32]. This is reflected in the ELBO term for the reverse diffusion process, which can be formulated as follows.

Proposition 3: The reverse diffusion process term in the ELBO can be rewritten as shown below (Proof: Supplementary Note 13):

$$\begin{aligned} &\sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)}\left[\mathbb{E}_{q_\phi(o,u|x_0)}[KL(q(x_{t-1}|x_t, x_0)|p(x_{t-1}|x_t, o, u)]\right] \\ &= \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)}\left[\mathbb{E}_{q_\phi(o,u|x_0)}\left[C_t - KL\left(p(x_{t-1}|x_t, o, u)||p(x_{t-1})\right)\right]\right] \end{aligned}$$

where C_t denotes the Kullback-Leibler (KL) divergence between the determined distribution $q(x_{t-1}|x_t, x_0)$ and the standard Gaussian distribution $p(x_{t-1}) := N(0, I)$.

Therefore, optimizing the model encourages the KL divergence $KL(p(x_{t-1}|x_t, o, u)||p(x_{t-1}))$ to approximate a constant C_t, effectively regulating the information content of x_{t-1}. The larger the KL divergence is, the more information x_{t-1} carries. By promoting $KL(p(x_{t-1}|x_t, o, u)||p(x_{t-1}))$ to approximate C_t, an information bottleneck effect is added to it and thus transferred to the latent factors o, u. Therefore, the diffusion model has a natural information bottleneck and is a good inductive bias for disentanglement representation purposes[32]. As different concepts to be disentangled may contain different amounts of information, the diffusion model can dynamically allocate this information to the latent concepts in the reverse diffusion step, where the information decreases as the number of time steps increases. This optimization objective is then similar to that of AnnealVAE[33].

Single-cell counterfactual generation via the do operator

CausCell performs interventions on individual cells by modifying their associated concepts, enabling the generation of counterfactual gene expression profiles. The causal disentanglement module of CausCell maps the concept representations of all N cells to a concept representation space Ω. In this space, each concept o_i has a representation for every cell, resulting in a total of N representations per concept. For each concept o_i, there exist m distinct concept labels labelled $a_1, a_2, \ldots, a_m$, each of which is associated with a subset of cells such that the total number of cells across all concept labels equals N.

To perform a concept intervention on a specific cell c that originally had a concept label of a_1 for concept o_i, we apply the do operator to set the concept to a new value a_k, which is denoted as $do(o_i = a_k)$. This intervention is implemented in three steps. (1) Randomly select a concept representation from the set of representations associated with the concept label a_k for concept o_i. (2) Replace the original concept representation of cell c with the selected representation. (3) Keep the representations of all other concepts $o_{j\neq i}$, including any unexplained concepts, unchanged for cell c. The postintervention distribution is formally represented as follows:

$$P\left(x_0 \middle| do(o_i = a_k), o_{j\neq i} = a_j^c, u = u^c\right)$$

where $o_{j\neq i} = a_j^c$ denotes that all other concepts o_j (for $j \neq i$) retain their original concept labels a_j^c, as in the original cell c. Utilizing these intervened concept representations, the generative diffusion module of CausCell generates a counterfactual gene expression profile $\hat{c}$ by sampling from the postintervention distribution:

$$\hat{c} \sim P\left(x_0 \middle| do(o_i = a_k), o_{j\neq i} = a_j^c, u = u^c\right)$$

This process allows us to simulate how the gene expression profile of cell c would appear under an intervention $do(o_i = a_k)$, isolating the causal effect of changing concept o_i from a_1 to a_k while controlling for all other concepts.

Data availability

All datasets utilized by CausCell for training, testing, and application examples were obtained from publicly accessible databases. Specifically, the Immun_atlas dataset was sourced from (https://www.tissueimmunecellatlas.org), the MERFISH_Brain dataset from (https://cellxgene.cziscience.com/collections/31937775-0602-4e52-a799-b6acdd2bac2e), and the Spatiotemporally_Liver dataset from (https://zenodo.org/records/7081863). The ICI_Response dataset was retrieved from the Gene Expression Omnibus (GEO) with accession number GSE123814, while the Limb_development dataset was accessed from ArrayExpress with accession ID E-MTAB-10514.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Grant No. T24250193, 32341008), the National Key Research and Development Program of China (Grant No. 2021YFF1201200, No. 2021YFF1200900), Shanghai Pilot Program for Basic Research, Shanghai Science and Technology Innovation Action Plan-Key Specialization in Computational Biology, Shanghai Shuguang Scholars Project, Shanghai Excellent Academic Leader Project, Shanghai Municipal Science and Technology Major Project (Grant No. 2021SHZDZX0100) and Fundamental Research Funds for the Central Universities. This work was partially funded by Microsoft Research Asia.

Author Contributions Statement

Qi Liu, Yicheng Gao and Kejing Dong designed the framework of this work. Dongsheng Li and Caihua Shan provided technical support. Yicheng Gao, Kejing Dong performed the analyses. Yicheng Gao, Kejing Dong, Caihua Shan and Qi Liu wrote the manuscript with the help of other authors. All authors read and approved the final manuscript.

Competing Interests Statement

The authors declare that they have no competing interests.

References

1. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics* **14**, 618-630 (2013).
2. Papalexi, E. & Satija, R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nature Reviews Immunology* **18**, 35-45 (2018).
3. Piran, Z., Cohen, N., Hoshen, Y. & Nitzan, M. Disentanglement of single-cell data with biolord. *Nature Biotechnology*, 1-6 (2024).
4. Bunne, C. et al. How to Build the Virtual Cell with Artificial Intelligence: Priorities and Opportunities. *arXiv preprint arXiv:2409.11654* (2024).
5. Higgins, I. et al. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230* (2018).
6. Bengio, Y., Courville, A. & Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* **35**, 1798-1828 (2013).
7. Buettner, F. et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology* **33**, 155-160 (2015).
8. Buettner, F., Pratanwanich, N., McCarthy, D.J., Marioni, J.C. & Stegle, O. f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome biology* **18**, 1-13 (2017).
9. Zhang, Z., Zhao, X., Bindra, M., Qiu, P. & Zhang, X. scDisInFact: disentangled learning for integration and prediction of multi-batch multi-condition single-cell RNA-sequencing data. *Nature Communications* **15**, 912 (2024).
10. Yu, H. & Welch, J.D. MichiGAN: sampling from disentangled representations of single-cell data using generative adversarial networks. *Genome biology* **22**, 158 (2021).
11. Lotfollahi, M. et al. Predicting cellular responses to complex perturbations in high□throughput screens. *Molecular systems biology* **19**, e11517 (2023).
12. Gabbay, A. & Hoshen, Y. Demystifying inter-class disentanglement. *arXiv preprint arXiv:1906.11796* (2019).
13. Pawlowski, N., Coelho de Castro, D. & Glocker, B. Deep structural causal models for tractable counterfactual inference. *Advances in neural information processing systems* **33**, 857-869 (2020).
14. Yang, M. et al. in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 9593-9602 (2021).

15. Croitoru, F.-A., Hondru, V., Ionescu, R.T. & Shah, M. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**, 10850-10869 (2023).
16. Domínguez Conde, C. et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022).
17. Allen, W.E., Blosser, T.R., Sullivan, Z.A., Dulac, C. & Zhuang, X. Molecular and spatial signatures of mouse brain aging at single-cell resolution. *Cell* **186**, 194-208. e118 (2023).
18. Afriat, A. et al. A spatiotemporally resolved single-cell atlas of the Plasmodium liver stage. *Nature* **611**, 563-569 (2022).
19. Yost, K.E. et al. Clonal replacement of tumor-specific T cells following PD-1 blockade. *Nature medicine* **25**, 1251-1259 (2019).
20. Zhang, B. et al. A human embryonic limb cell atlas resolved in space and time. *Nature*, 1-11 (2023).
21. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature methods* **15**, 1053-1058 (2018).
22. Lotfollahi, M., Wolf, F.A. & Theis, F.J. scGen predicts single-cell perturbation responses. *Nature methods* **16**, 715-721 (2019).
23. Kim, H. & Mnih, A. in International conference on machine learning 2649-2658 (PMLR, 2018).
24. Higgins, I. et al. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)* **3** (2017).
25. Ashburner, M. et al. Gene ontology: tool for the unification of biology. *Nature genetics* **25**, 25-29 (2000).
26. Hou, J. et al. Antibody-mediated targeting of human microglial leukocyte Ig-like receptor B4 attenuates amyloid pathology in a mouse model. *Science Translational Medicine* **16**, eadj9052 (2024).
27. Zhou, J. et al. LilrB3 is a putative cell surface receptor of APOE4. *Cell Research* **33**, 116-130 (2023).
28. Song, Y. & Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* **32** (2019).
29. Tang, W. et al. A general single-cell analysis framework via conditional diffusion generative models. *bioRxiv*, 2023.2010. 2013.562243 (2023).
30. Hou, R., Chang, H., Ma, B., Shan, S. & Chen, X. Cross attention network for few-shot classification. *Advances in neural information processing systems* **32** (2019).
31. Savani, Y., White, C. & Govindarajulu, N.S. Intra-processing methods for debiasing neural networks. *Advances in neural information processing systems* **33**, 2798-2810 (2020).
32. Yang, T., Lan, C. & Lu, Y. Diffusion Model with Cross Attention as an Inductive Bias for Disentanglement. *arXiv preprint arXiv:2402.09712* (2024).
33. Burgess, C.P. et al. Understanding disentangling in $\beta $-VAE. *arXiv preprint arXiv:1804.03599* (2018).