OXFORD

*Briefings in Bioinformatics*, 2024, **25(6)**, bbae586
https://doi.org/10.1093/bib/bbae586
**Problem Solving Protocol**

# Robust self-supervised learning strategy to tackle the inherent sparsity in single-cell RNA-seq data

Sejin Park[1] and Hyunju Lee[1,2,*]

[1]School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, 61005, Gwangju, South Korea
[2]Artificial Intelligence Graduate School, Gwangju Institute of Science and Technology, 61005, Gwangju, South Korea

*Corresponding author: E-mail: hyunjulee@gist.ac.kr

## Abstract

Single-cell RNA sequencing (scRNA-seq) is a powerful tool for elucidating cellular heterogeneity and tissue function in various biological contexts. However, the sparsity in scRNA-seq data limits the accuracy of cell type annotation and transcriptomic analysis due to information loss. To address this limitation, we present scRobust, a robust self-supervised learning strategy to tackle the inherent sparsity of scRNA-seq data. Built upon the Transformer architecture, scRobust employs a novel self-supervised learning strategy comprising contrastive learning and gene expression prediction tasks. We demonstrated the effectiveness of scRobust using nine benchmarks, additional dropout scenarios, and combined datasets. scRobust outperformed recent methods in cell-type annotation tasks and generated cell embeddings that capture multi-faceted clustering information (e.g. cell types and HbA1c levels). In addition, cell embeddings of scRobust were useful for detecting specific marker genes related to drug tolerance stages. Furthermore, when we applied scRobust to scATAC-seq data, high-quality cell embedding vectors were generated. These results demonstrate the representational power of scRobust.

**Keywords**: contrastive learning; self-supervised learning; cell type annotation; marker gene discovery

## Introduction

Single-cell RNA-sequencing (scRNA-seq) has gained prominence for its ability to reveal the distinct features of tissues and organisms. Although scRNA-seq allows high-resolution analysis at the individual cell level, it does not inherently provide cell-type labels, which limits transcriptomic analysis. Consequently, cell type annotation is the first step in scRNA-seq, for which several computational methods, like Seurat [1] and JAGLRR [2], have been developed, but do not address the issue of sparsity in scRNA-seq data. The depth of coverage in scRNA-seq is a trade-off between the ability to detect a large number of cells simultaneously. For example, platforms like 10X Genomics Chromium [3] can detect many cells relatively inexpensively but suffer from significant dropouts. In contrast, platforms such as Smart-seq2 [4] provide high depth of coverage but tend to detect fewer cells at a high cost. Therefore, addressing this issue is crucial for achieving more robust analyses and accurate biological insights.

Several Transformer-based models tailored for scRNA-seq [5–7] have been introduced. Because the Transformer was originally designed for natural language processing (NLP) tasks, genes and their expression values require transformation to fit into NLP models. For instance, scGPT [8] transforms genes into words and categorizes gene expression values into bins based on their ranges. CIForm [7] divides lengthy cell-based gene expression vectors into sub-vectors for projection into an embedding space akin to ViT [9]. TOSICA [6] generates different gene expression vectors by masking original vectors, and then projects them into an embedding space. scFoundation [10] uses gene expression embedding vectors and combines gene embedding vectors with gene expression vectors.

Self-supervised learning (SSL) is required for initializing large language models. Contrastive learning has been used in various tasks and showed remarkable improvement in downstream tasks [11–14]. In contrastive learning, data augmentation creates pseudo-objects with local information derived from the original object containing global information. Designing high-quality data augmentation approaches for scRNA-seq is more challenging than in computer vision, due to the extensive number of genes. In computer vision, a given image contains all its information, allowing images to be augmented from complete information. However, in scRNA-seq, a subset, e.g. 2,000 highly variable genes (HVGs), is more often employed than entire gene data because of the large dimension. Therefore, data augmentation from the subset will suffer from severe incomplete information, and it is challenging to leverage global information of scRNA-seq effectively.

CLEAR [15], Concerto [14], and Cake [16] apply contrastive learning to scRNA-seq data. Although CLEAR creates augmented embeddings using various augmentation techniques across all genes, the augmented scRNA-seq data exhibits even greater sparsity than the original data. In contrast, Concerto and Cake utilize 2,000 HVGs and generate two embedding vectors for each target cell using two different encoders. Specifically, Cake employs a K-nearest neighbor search algorithm [17] to create pseudo labels and assigns the same positive label to cells with identical pseudo labels. On the other hand, Concerto augments samples by

**Received:** July 13, 2024. **Revised:** September 26, 2024. **Accepted:** October 31, 2024

applying different encoders to the same cell; the positive pair consists of embedding vectors of the same cell generated from two different encoders. As a result, these augmentation techniques using contrastive learning are still limited to a subset of all genes and are ineffective at learning global information and mitigating data sparsity.

This study presents and validates scRobust, a robust self-supervised learning strategy designed to address the inherent sparsity of scRNA-seq data. scRobust pre-trains a Transformer encoder through contrastive learning and gene expression prediction. Unlike previous foundation models [8, 10], we introduce a novel self-supervised learning strategy specifically tailored for scRNA-seq data, incorporating a novel cell augmentation technique to capture information across all genes. After pre-training, the encoder is fine-tuned for cell-type annotation tasks using highly unique genes. We demonstrate that scRobust outperforms benchmark methods across most datasets and generates cell embeddings that contain both cell-type and sample-specific information. Additionally, we showed that our approach is also applicable to scATAC-seq data.

# Results

## Overview of scRobust

The framework of scRobust is divided into the pre-training (Fig. 1a–c) and downstream phases (Fig. 1d and e). For contrastive learning, we developed **cell augmentation**, which generates high-quality data augmentation of scRNA-seq data (Fig. 1a). In cell augmentation, scRobust generates diverse cell embeddings for a target cell from random gene sets without dropout. The cell augmentation encoder offers two primary benefits: firstly, it generates varied and distinct cell-embedding vectors by utilizing numerous combinations of partial local information (small subsets of whole genes); secondly, the encoder can learn all genes with non-zero expression values. Therefore, the encoder can access global information (whole genes) and effectively address scRNA-seq data sparsity.

In the contrastive learning stage (Fig. 1b), the encoder creates various local cell embeddings for each sample via cell augmentation. A simple classifier identifies the correct pairs of embedding vectors originating from the same cell. These embeddings attract and repel those from the same and different cells, respectively, and each cell establishes a unique territory within the cell embedding space (see "Contrastive learning" in the Methods). Therefore, scRobust can map any local cell embeddings into a distinct area within the cell embedding space, effectively aligning local with global cell embeddings. During gene expression prediction (Fig. 1c), the encoder predicts the expression of certain genes via the dot product between a given local cell embedding and target gene embeddings. In Figs S1 and S2, we verified that the losses of contrastive learning and gene expression prediction tasks were decreased in the training and test sets. This indicates that our model can effectively extract cell-specific information representing a given cell using a small number of different genes. The details are provided in the "Pre-training results" of the supplementary file.

After pre-training, the gene embeddings and encoder were fine-tuned for cell type annotation. In the downstream task (Fig. 1d and e), we used a larger number of genes relative to the random genes used in pre-training, and the input gene sets consisted of highly unique genes, which differ between cells. In this context, highly unique genes refer to genes rarely expressed within the population of a given dataset. As we used highly unique but non-zero read count genes as the input, each cell embedding was generated with different non-zero genes. Thus, this approach effectively mitigates scRNA-seq data sparsity (see "Unique gene selection" in the Methods; the benefits of using highly unique genes are elaborated on in "Impact of unique genes" of the supplementary file and Fig. S3).

## Datasets

To assess the performance of scRobust in cell type annotation across different protocols and tissues, we utilized nine benchmark datasets: Baron Human [18], Muraro [19], Segerstolpe [20], Xin [21], TM [22], Zheng 68K [3], Zheng sorted [3], MacParland [23], and Baron Mouse [18]. Additionally, to thoroughly test the performance of scRobust against the inherent sparsity of scRNA-seq data, we generated corrupted datasets with severe data dropout. Specifically, we added artificial dropouts of 30% and 50%, converting the read counts of the corresponding genes with non-zero values to zero in each cell. In these corrupted scenarios, we used the models pre-trained on these datasets with additional dropouts. The results represent the averages from 5×5 cross-validation in the Baron Human [18], Muraro [19], Segerstolpe [20], Xin [21], MacParland [23], and Baron Mouse datasets [18]. For the TM [22], Zheng 68K [3], and Zheng sorted datasets [3], five-fold cross-validation was employed due to the larger number of cells. The details are in "Description of datasets" of the supplementary file and Table S1.

## Cell type annotation in nine datasets

Figure 2a shows the macro F1 and accuracy scores for scRobust and seven benchmark methods across all datasets and dropout scenarios (detailed in the "Benchmark methods" section of the supplementary file). scRobust achieved the highest F1 scores in eight out of nine datasets, with the exception of the Xin dataset [21]. Figure 2b presents heatmaps of class-wise accuracy scores for the Zheng 68K, Muraro, Baron Mouse, and Segerstolpe datasets, comparing scRobust to Concerto [14], CIForm [7], and TOSICA [6]. Figure S4 shows the cell counts for each cell type across the nine datasets, along with counts of rare cell types within each dataset. From Figs 2b and S4, it was observed that scRobust significantly outperformed the benchmark methods in identifying rare cell types. For instance, in the Zheng 68K dataset, scRobust achieved an accuracy of 0.28 for CD4+ T Helper 2 cells, while the other methods had accuracies below 0.10. In the Muraro dataset, scRobust achieved an accuracy of 1.0 for epsilon cells, while the other methods scored zero. Similar trends were observed in the Baron Mouse dataset for T, B, and Schwann cells, and in the Segerstolpe dataset for MHC class II and epsilon cells. These results demonstrate that scRobust excels in classifying rare cell types.

In the experiments involving additional dropout, scRobust consistently achieved the highest performance across all cases and datasets. With an additional 30% dropout, scRobust demonstrated superior performance compared to the benchmark methods without additional dropout in the TM, Zheng sorted, Segerstolpe, and Baron Mouse datasets (Fig. 3a). Even in the same datasets, the performance of the second-best methods was similar to that of scRobust with 50% additional dropout. With the exception of CLEAR [15], most methods showed notable declines in performance with additional dropout. However, scRobust was significantly less affected by additional dropout, underscoring its performance against scRNA-seq data sparsity.

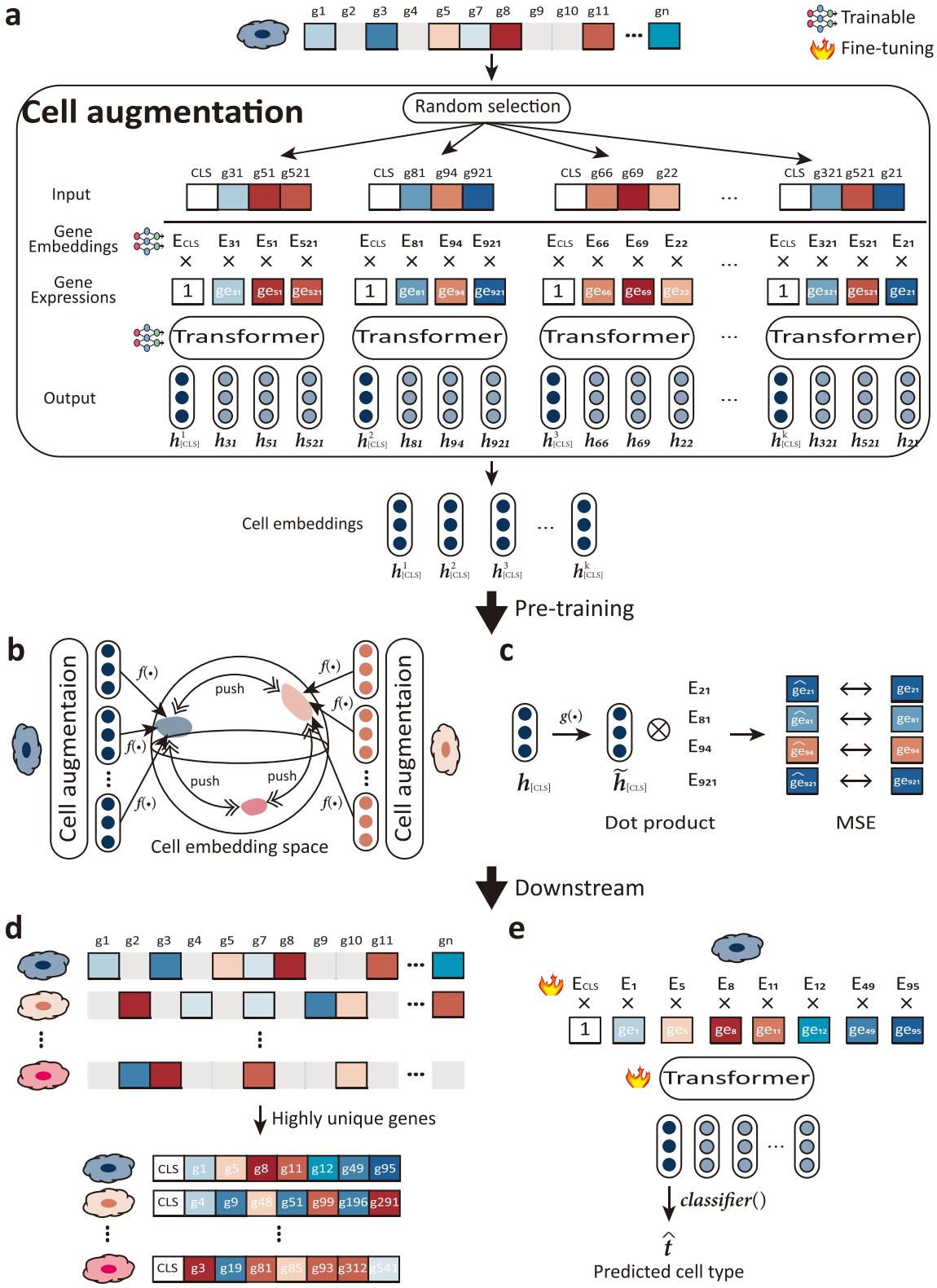


Figure 1. **Overview of pre-training and downstream tasks in scRobust. (a)** Cell augmentation generates various cell embeddings using different randomly selected gene sets from a given cell. For clarity, Transformer was drawn multiple times although a single Transformer was used. **(b)** Contrastive learning, where local cell embeddings from the same and different cells attract and repel each other, respectively. **(c)** Gene expression prediction task, where an arbitrary local cell embedding is used to predict the expression of randomly selected genes, which differ from the genes used to generate the local cell embedding. **(d)** Gene selection for a downstream task, where exclusively expressed genes in a target cell (i.e. highly unique genes) are preferentially selected. **(e)** Cell type annotation for the downstream task, using highly unique genes and the pre-trained encoder.

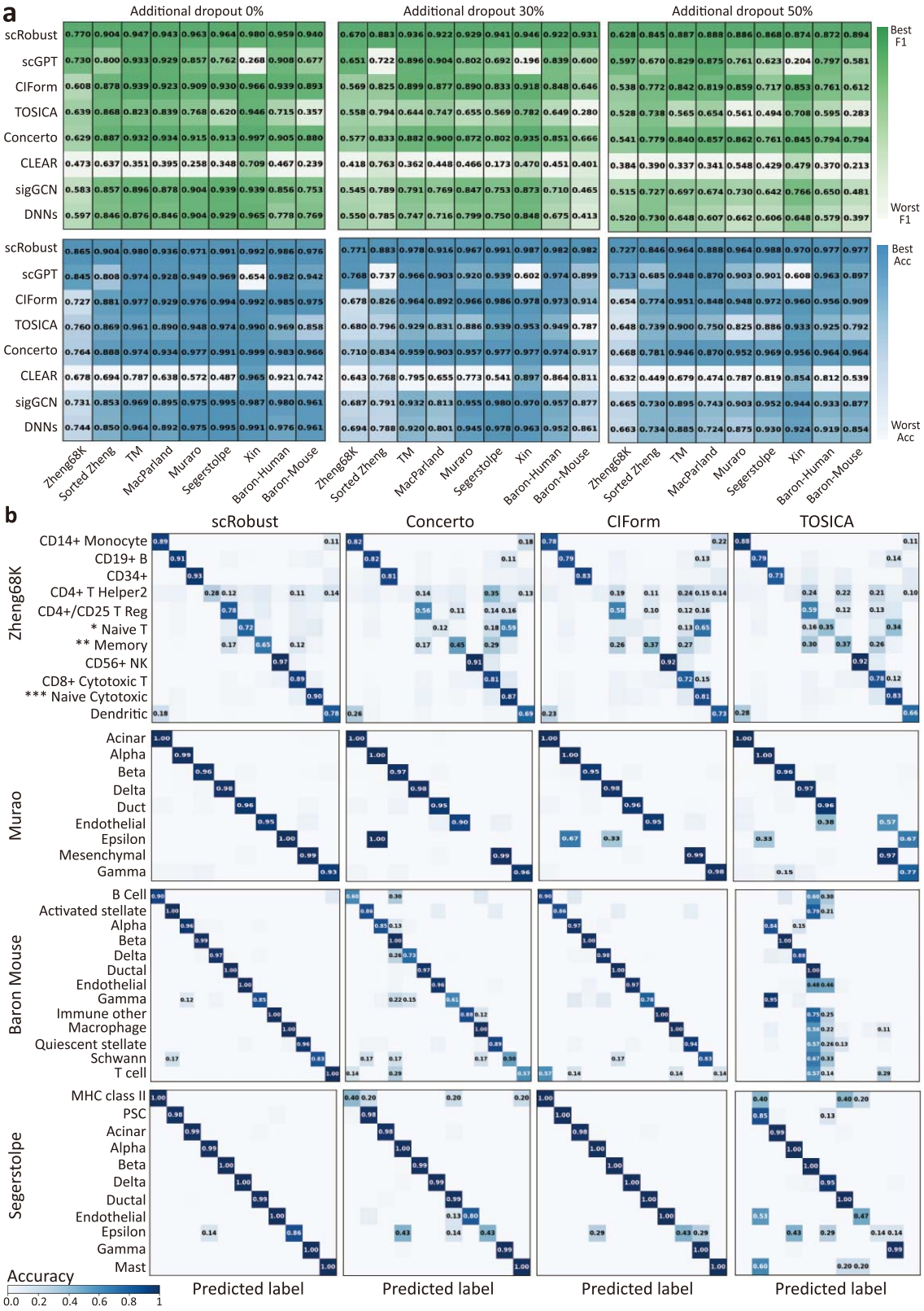


Figure 2. **Cell type annotation results in intra-dataset cross-validation. a** Comparison of the F1 and accuracy scores at 0%, 30%, and 50% additional dropout, illustrating the performance of scRobust under varying degrees of data sparsity. **b** Heatmaps based on confusion matrices for scRobust, Concerto, CIForm, and TOSICA. These matrices display the accuracy of each cell type in the Zheng 68K, Muraro, Baron Mouse, and Segerstolpe datasets. *, **, and *** denote CD4+/CD45RA+/CD25-, CD4+/CD45RO+, and CD8+/CD45RA+, respectively, highlighting specific cell types within the datasets.

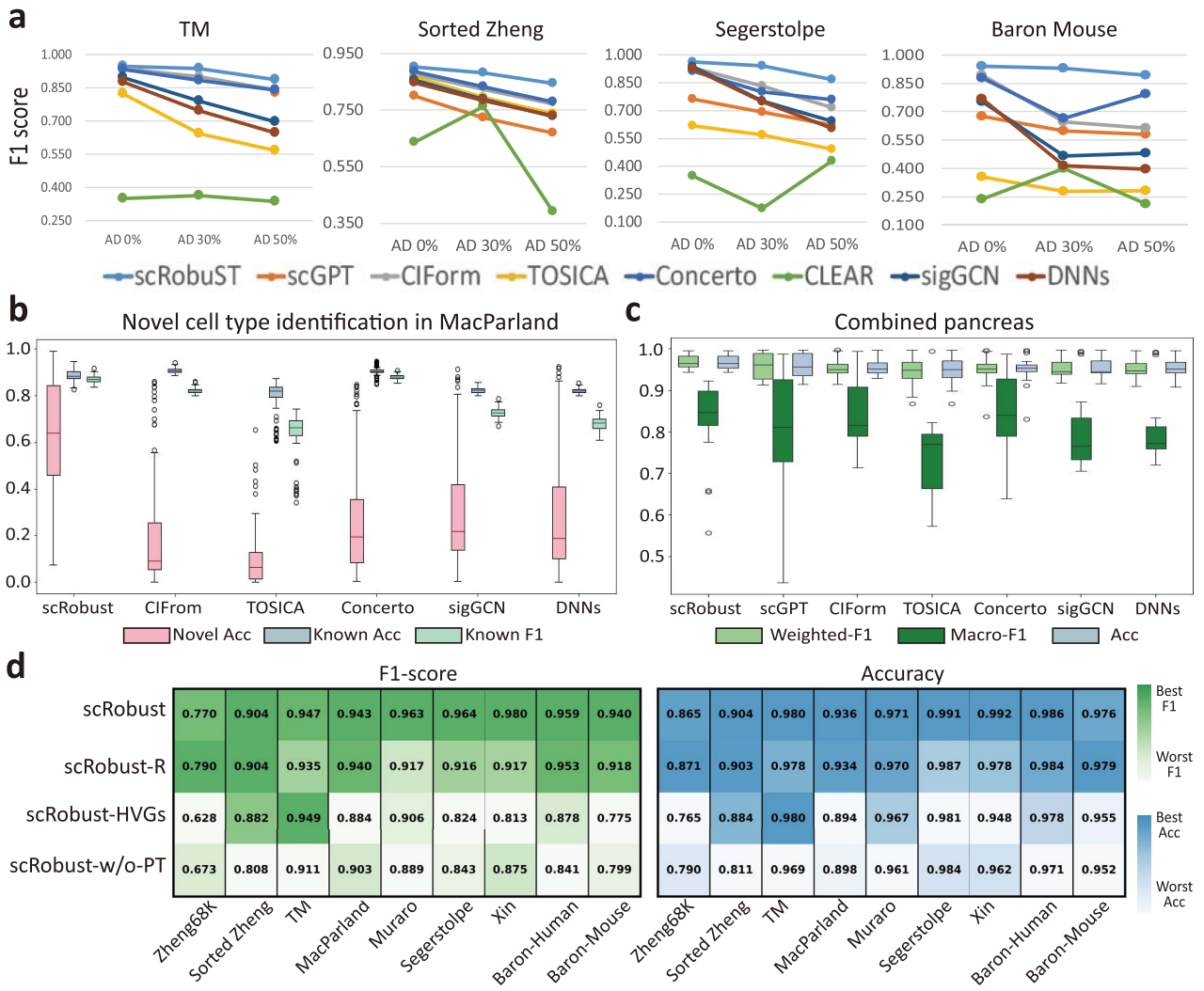


Figure 3. **Performance of scRobust in various environments. (a)** Line charts depicting F1 scores under various additional dropout scenarios, with "AD" signifying "additional dropout." The chart illustrates how scRobust's performance varies with increasing levels of dropout. **(b)** Box plot of scRobust's performance in predicting novel cell types, highlighting its ability to handle unseen cell types. **(c)** Box plot demonstrating the performance of scRobust across different sequencing platforms, indicating its adaptability to various data sources. **(d)** Results from an ablation study conducted on scRobust, illustrating the impact of different factors. Here, "R" refers to the use of random genes, "HVGs" indicates the use of highly variable genes, and "w/o-PT" represents scenarios without pre-training.

## Detection of novel cell types

As unknown cell types are frequently encountered in real-world scenarios, the efficient identification of novel cell types is crucial in practice. Accordingly, we assessed average model performance when encountering novel cell types across five runs using the MacParland dataset [23]. In this task, we used a frozen pre-trained scRobust to avoid overfitting (details in "Discovery of novel cell type" of the supplementary file).

Methods that perform well with known cell types often exhibit poorer performance when encountering novel cell types. Consequently, simpler methods are generally outperformed by more complex methods for unknown (novel) cell types. Figure 3b illustrates this phenomenon, where the simpler methods sigGCN and DNNs outperform the more complex models CIForm, TOSICA, and Concerto. The F1 or accuracy scores of CIForm and Concerto were 0.819 or 0.907 and 0.907 or 0.905, respectively, whereas their accuracy scores for novel cell types were 0.202 and 0.265, respectively. However, scRobust outperformed the other methods for both known (F1 score = 0.870, accuracy = 0.888) and novel cell types (accuracy = 0.619). Considering that a frozen pre-trained scRobust was used in this task, our findings suggest that the encoder was effectively trained to extract cell-type information during the pre-training phase.

## Cell type annotation across different sequencing platforms

Cell type annotation across diverse datasets is an essential task in real-world applications, as single-cell datasets often originate from varied batches and platforms. To assess the performance in annotation across different sequencing platforms, we created a combined human pancreas dataset using the Baron Human [18], Muraro [19], Segerstolpe [20], and Xin datasets [21]. Because some cell types do not exist in a test dataset, the macro-F1 score could not properly evaluate models when predicting cell types not existing in a test dataset. Thus, we evaluated models using accuracy, the macro-, and weighted-F1 scores.

Figure 3c and Table S2 show that scRobust outperformed benchmark methods in weighted-F1 (0.968 ± 0.016) and accuracy (0.967 ± 0.017), but not in macro-F1 (0.827 ± 0.097). Among benchmarks, CIForm had the highest macro-F1 (0.843 ± 0.091)

but ranked third in weighted-F1 (0.954 ± 0.025) and accuracy (0.956 ± 0.023). scGPT was second in weighted-F1 (0.958 ± 0.029) and accuracy (0.958 ± 0.028) but had a lower macro-F1 (0.803 ± 0.137). The macro-F1 significantly decreases with incorrect predictions of cell types not in the test set. For example, in the Xin dataset with four cell types, some methods showed lower macro-F1 than weighted-F1. scRobust's narrow, high-positioned box plots in Fig. 3c and its low standard deviations in weighted-F1 and accuracy indicate its high and stable performance across different datasets and sequencing platforms.

## Performance in complex tissues

To verify whether scRobust can extract meaningful biological information from complex tissues, we tested our model on datasets from the cerebellum, multiple sclerosis (MS), cortex, kidney, heart, and cross-tissue immune cells [24–29] (Detailed in the "Complex tissue datasets" section of the supplementary file and in Table S3). The cerebellum, MS, cortex, kidney, heart, and immune datasets contain 59, 18, 19, 26, 60, and 45 cell types, respectively. For the cortex dataset [26], we used 19 subclasses that are matched with subtypes in [30].

As shown in Fig. S5, scRobust achieved the best performance in the MS, kidney, heart, and cross-tissue immune cell datasets, whereas CIForm outperformed scRobust in the cerebellum and cortex. Although the performances between scRobust and CIForm were similar in the cerebellum dataset, scRobust performed the worst in the cortex dataset. Thus, to investigate the reason, we generated t-SNE plots of GABAergic cell embeddings from scRobust and principle component analysis (PCA). In Fig. S6, GABAergic PAX6 cells were found to be mixed with some GABAergic VIP cells only in the scRobust, whereas these cell types were clearly distinguished in the PCA. However, PAX6 expression was present only in a subset of GABAergic PAX6 cells, and GABAergic VIP cell in the mixed area rarely had gene expression value of VIP. Instead, the cells in the mixed region showed significant expression levels of CRN1. To explore the biological states of the isolated versus mixed GABAergic VIP cells, we conducted an enrichment test (GSEA) [31]. The GABAergic VIP cells in the mixed area had the low normalized enrichment scores (NES) for `GOBP_CENTRAL_NERVOUS_SYSTEM_NEURON_DEVELOPMENT` and `GOBP_CENTRAL_NERVOUS_SYSTEM_NEURON_DIFFERENTIATION` and high NES for `REACTOME_GABA_RECEPTOR_ACTIVATION` (Table S4). Considering the expression levels of marker genes and the enrichment test results, it is possible to classify cells in the mixed area as another sub-type. These results indicate that brain cells with a variety of characteristics could be classified in several different ways.

## Ablation study

We designed three variations of scRobust for the ablation study: scRobust-*w/o*-PT is scRobust without pre-training; scRobust-RGs uses random genes instead of unique genes; scRobust-HVGs uses HVGs instead of unique genes (details in "Ablation study design" of the supplementary file). Figure 3d shows the results of the ablation study. As expected, scRobust-*w/o*-PT performed worse than other scRobust variations in most of the datasets. Interestingly, the performance of scRobust-HVGs was similar to that of scRobust-*w/o*-PT in some datasets. Despite using a pre-trained model, employing the same genes for all cells seems to limit the extraction of unique features from each cell. Conversely, using individual random genes (scRobust-RGs) achieved the second-highest performance across most datasets. This suggests that the pre-trained model can extract global information from randomly selected genes. Nevertheless, utilizing highly unique genes consistently ensures more stable and superior performance compared to random genes in cell type annotation tasks.

## Pathways in the cell embedding space

Considering that pathways consist of specific gene sets, scRobust can generate pathway-related embedding vectors using the genes from a given pathway. For this process, we define two types of vectors: pathway-informed cell embedding vectors and pathway embedding vectors (see "pathway embedding vector" in the Methods). The pathway-informed cell embedding vector is generated using a pathway gene set and gene expression values from an input cell. Similarly, the pathway embedding vector also utilizes genes from the pathway, but not scaled by gene expression values. In the cell embedding space, pathway embedding vectors represent the pathways themselves, whereas pathway-informed cell embedding vectors encapsulate information from the given pathway for individual input cells.

Some cell types can be characterized by specific pathways; e.g. Mauro *et al.* [19] identified pathways that define different cell types in the pancreas, such as alpha, beta, and gamma cells (Table S5). To visualize the relationship between pathways and cell embedding vectors, we drew t-SNE plots of alpha, beta, gamma, and delta cells, along with their corresponding pathway vectors in Segerstolpe [20], Xin [21], Muraro [19], and Baron human [18]. As shown in Fig. 4a, the pathway vectors were clustered near the cell embedding vectors of the same cell type. Additionally, we can predict cell types by evaluating the closeness between pathway-informed cell embedding vectors and their corresponding cell-type pathways (see "pathway embedding vector" in the Methods). This approach yielded reasonable results across four pancreas datasets: Acc = 0.891, F1 = 0.806 for Xin; Acc = 0.806, F1 = 0.701 for Baron human; Acc = 0.836, F1 = 0.805 for Segerstolpe; and Acc = 0.905, F1 = 0.840 for Muraro. Notably, this approach showed lower performance only for ductal cells (Acc = 0.349 in Baron, Acc = 0.498 in Segerstolpe, and Acc = 0.473 in Muraro), suggesting the potential for improvement when utilizing other ductal cell pathways. It is important to emphasize that this approach is unsupervised, and can be applied if well-defined cell-type pathways exist, providing a promising alternative for cell-type annotation.

## Impact of the pre-training tasks

The fundamental principle of our pre-training strategy is to assimilate global information from all genes. By considering all genes and their interrelationships, scRobust may capture both the shared features of various cell types and the unique characteristics of individual cells. In this experiment, we used the Segerstolpe dataset [20], which contains the transcriptomes of human islet cells for analyzing diabetes characteristics at the single-cell level. The dataset contains cell type information for each cell and hemoglobin A1c (HbA1c) values, used to assess glucose control. Given that diabetes is a complex disease impacting various bodily systems, HbA1c serves as an effective indicator for capturing the individual characteristics of cells. In practice, it may be useful to simultaneously examine cell-type and sample-specific information such as the HbA1c value. Accordingly, we compared the cell embeddings of scRobust with those generated by Concerto and 2,000 HVGs.

We examined the clustering of cells based on cell type and HbA1c values using scRobust. The experiment focused on $\alpha, \beta, \delta$, and $\gamma$ cell types, which are directly linked to diabetes, and generated t-SNE plots. In Fig. 4b, the clusters in scRobust and HVGs are well distinguished by cell type (dot lines), unlike in Concerto.

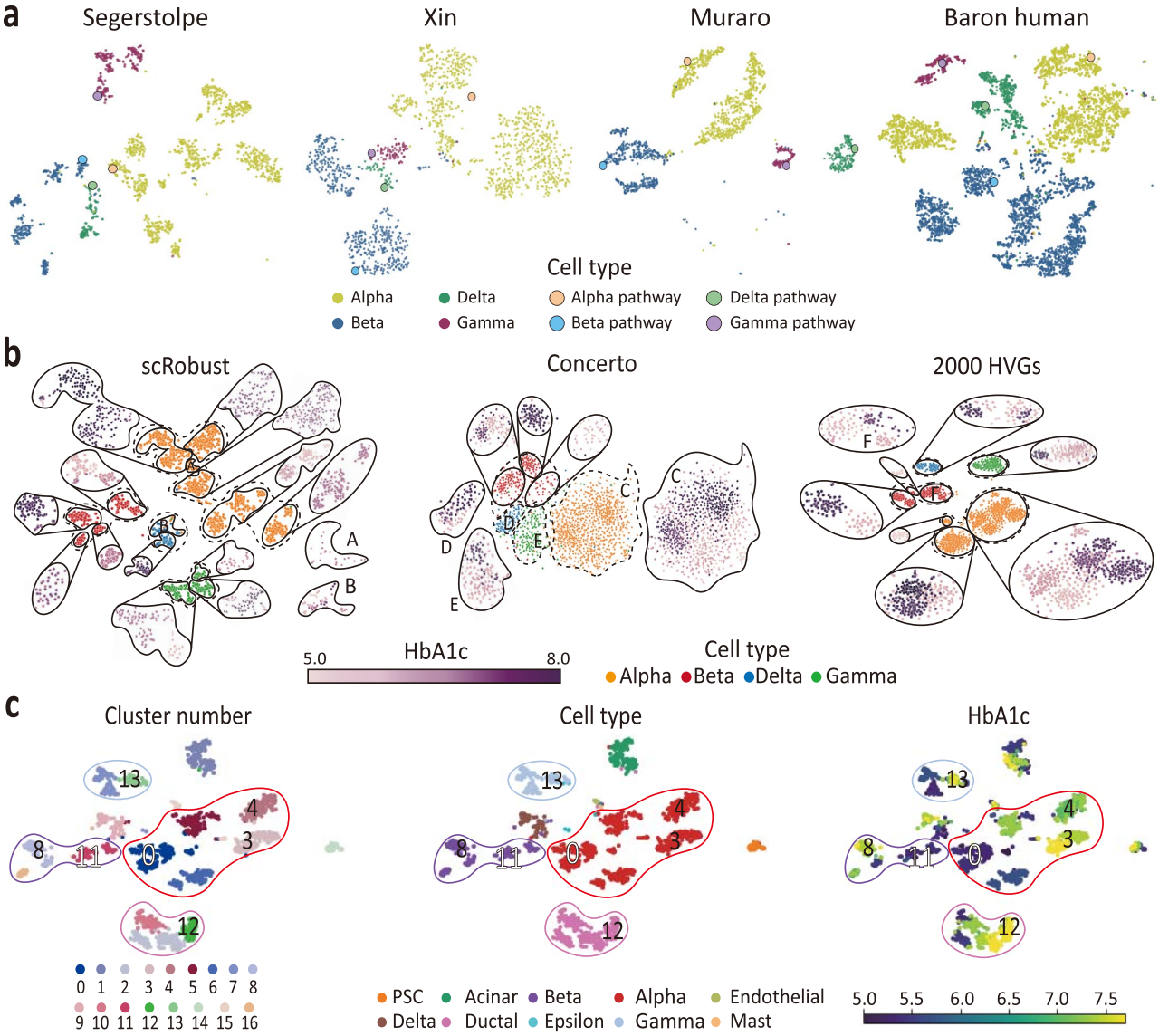


Figure 4. **t-SNE plots of cell-embedding vectors without fine-tuning. (a)** t-SNE plots showing alpha, beta, gamma, and delta cells, along with their corresponding pathway embedding vectors for the Segerstolpe, Xin, Muraro, and Baron human datasets. **(b)** Comparative analysis of scRobust, Concerto, and 2000 HVGs in the Segerstolpe dataset, where the color of the points represents HbA1c levels. Dotted lines distinguish between cell type groups, while solid lines separate clusters within those groups. **(c)** Cell embeddings generated by scRobust, labeled by (left) cluster numbers, (center) cell types, and (right) HbA1c levels for all cells in the Segerstolpe dataset. The numbers in clusters are cluster numbers generated by the Leiden algorithm, and clusters enriched with patient and normal samples are numbered with black and white colors, respectively. The colored line indicates the cell cluster, and these colors are the same as the labels' colors in the center.

Furthermore, several islands are observed within the same cell type clusters for scRobust, in contrast to the more uniform clustering seen with HVGs. The island formation in scRobust seemed to be associated with the comparable HbA1c values between cells in the small islands (solid lines). In contrast, the cells in the other methods were mixed in terms of HbA1c values. To ensure that the clusters generated by scRobust were influenced by HbA1c (diabetes) levels rather than batch effects, we created t-SNE plots labeled by sample IDs and cluster numbers assigned by Leiden algorithm [32] (Fig. S7 and Table S6) and conducted a gene set enrichment analysis (GSEA) comparing clusters with high and low HbA1c values. The results showed that most clusters comprised different sample IDs, and many pathways related to diabetes and insulin were identified between normal and diabetic cell clusters (Tables S7–13).

As such, scRobust tends to generate multiple clusters even within the same cell types. Thus, we furthermore investigated whether different clusters within the same cell type may reflect distinct biological pathways. To explore this, we generated t-SNE plots of the Segerstolpe dataset [20] for all cell types, distinguishing the cells by cluster numbers identified by the Leiden algorithm, their cell types, and HbA1c values (Fig. 4c). Specifically, we observed that $\alpha, \beta, \delta$ and ductal cells formed five, three, two, and three clusters, respectively. Then, we conducted GSEA [31] between target clusters and the remaining clusters (Tables S14–20). In the GSEA of $\alpha$ cell clusters, the `MURARO_PANCREAS_ALPHA_CELL` pathway was enriched in both clusters #0 and #3, but normalized enrichment scores (NES) of clusters #0 and #3 were -2.52 and 2.05, respectively. It is remarkable that most cells in cluster #3 were from diabetic patients, whereas cluster #0 primarily consisted of normal cells. Additionally, the diabetes-related pathway, `GSE9006_HEALTHY_VS_TYPE_2_DIABETES_PBMC_AT_DX_UP`, was enriched in cluster #4, where most cells exhibited high HbA1c

values. In $\beta$ cell clusters, $\beta$ cell-related pathways (MURARO_PANCREAS_BETA_CELL and VANGURP_PANCREATIC_BETA_CELL) were enriched in both clusters #11 (normal) and #8 (patient). Like $\alpha$ cell clusters, their NESs were low in the normal cluster and high in the patient cluster. Similarly, cluster #13, which consists of gamma cells from patients, showed high NES for the MURARO_PANCREAS_PANCREATIC_POLYPEPTIDE_CELL and VANGURP_PANCREATIC_GAMMA_CELL pathways. We consistently observed that patient-derived clusters had high NESs for the MURARO and VANGURP pathways, while normal clusters had low NESs for these pathways. Additionally, the ductal cell cluster containing patient cells had low NESs for insulin secretion related pathways. These results demonstrate that scRobust not only identifies cell type information but also detects detailed biological states by capturing functionally related gene sets (i.e. pathways) even in sparse data.

Additionally, to verify how well scRobust classifies detailed information, we generated t-SNE plots for the Baron Human [18] and sorted Zheng datasets [3](Fig. S8a). Although the t-SNE plots of scRobust showed various islands, the cell embeddings effectively distinguished cell types. For instance, in the Baron Human dataset, scRobust successfully separated activated and quiescent stellate cells and formed a distinct cluster for Schwann cells. In contrast, these cell types were not clearly distinguished and were mixed in the plots for the other methods. In the sorted Zheng dataset, the cell embeddings of scRobust were more clearly arranged by cell type, whereas the cell embeddings of the other methods were largely intermixed.

We hypothesized that cell embeddings generated from different layers of the encoder may contain distinct types of information. Accordingly, we used the combined pancreas dataset (Xin, Muraro, Segerstolpe, and Baron Human), which contains cell-type information and batch effects, and pre-trained scRobust with two layers. Figure S8b shows the t-SNE plots of cell embeddings from the first and second layers of scRobust, Concerto, and 2,000 HVGs. The cell embeddings for Concerto and HVGs were poorly clustered by cell type across datasets, indicating that cell-type information is overshadowed by batch effects. In contrast, cell embeddings from the first layer of scRobust were predominantly clustered according to cell type, with distinct dataset information classified within the larger clusters. Conversely, the cell embeddings of the second layer were predominantly clustered by dataset, with cell type-specific clusters manifesting within each dataset cluster. Therefore, each layer of scRobust targets different information, suggesting that cell embeddings from different layers may be suitable for varied analytical purposes. Notably, no labels, such as dataset platforms, were used during the pre-training stage.

## Discovery of marker genes

The discovery of marker genes is an important topic in scRNA-seq annotation. scRobust can recommend important genes as marker genes using the attention scores of its encoder (details in "Marker gene selection" of the supplementary file). Tables S21–S29 show the 10 selected genes for each cell type in all datasets. Among these selected genes, most are marker genes with literature evidence on their relationship between the gene and its cell type. We generated heatmaps to visually verify the performance of scRobust in recommending marker genes (Fig. 5a and Fig. S9). scRobust generated distinct clusters for most cell types across all datasets based on the selected genes.

Marker genes are often overexpressed in the target cell type; therefore, marker gene information may influence classification accuracy. We hypothesized that scRobust may identify more marker genes for cell types where it performed well. In the Zheng 68K dataset [3] (Fig. 2b), scRobust achieved 97% accuracy for CD56+ NK cells, identifying nine out of 10 selected genes as marker genes. For CD14+ monocytes and CD19+ B cells, it achieved 89% and 91% accuracy, with eight and seven marker genes, respectively. In the MacParland dataset [23] (Fig. S10), accuracy scores for five cell types ranged from 98% to 100%, with over 80% of selected genes being marker genes. In the Baron Mouse dataset [18], scRobust identified all selected genes as marker genes for T cells (accuracy 100%), unlike CIForm and Concerto, which had lower performances (14% and 57%, respectively) due to using 2,000 HVGs without marker genes (Fig. 2b).

scRobust performed well in distinguishing between cells that share the same general cell type but have different sub-cell types or cell states, such as inflammatory and non-inflammatory macrophages (Fig. 2b and Fig. S10). To investigate the performance of scRobust in extracting general cell-type and sub-cell-type information, we analyzed the recommended genes using scRobust in the MacParland, Baron Human, and Baron Mouse datasets, which contain sub-cell type relationship information. Figure 5b shows the differential expression of the marker genes identified by scRobust across different sub-cell types or cell states. Marker genes selected in each subtype, such as S100A12, CD5L, and MARCO, were exclusively expressed, demonstrating the effective extraction of detailed information. In contrast, FCER1G was overexpressed in both sub-cell types because it was chosen as a marker gene for both inflammatory and non-inflammatory macrophages.

In the MacParland dataset, the marker genes reflect the roles of each subtype. S100A12, a protein expressed in myeloid cells, acts as a proinflammatory alarm signal and is associated with the inflammatory response [33]. Conversely, MARCO is expressed on tumor-associated macrophages that adopt an immunosuppressive phenotype, counteracting inflammation within the tumor microenvironment [34]. Therefore, we verified that scRobust can extract not only common but also detailed sub-type information in marker gene discovery.

## Identifying drug tolerance stages and marker genes

Aissa *et al.* [35] investigated changes in populations of the non-small-cell lung cancer (NSCLC) cell line PC9 with varying degrees of erlotinib tolerance over days. They observed that the earliest drug-tolerant persisters (DTPs, observed on Days 2 and 4) and the drug-tolerant expanded persisters (DTEPs, observed on Days 9 and 11) exhibit tolerance to significantly higher concentrations of erlotinib compared to the untreated original PC9 cells (Day 0). They presented UMAP representations of PC9 cells by treatment days for an integrated dataset (Days 0, 1, 2, 4, 9, and 11) without correcting for batch effects. However, it is crucial to remove batch effects because the dataset itself serves as the label, making it difficult to determine whether clustering is due to batch effects or an ability of single-cell representation to distinguish transcriptional changes associated with drug tolerance. After removing batch effects using ComBat [36], we found that cells represented with 1,000 HVGs were mixed regardless of treatment days (Fig. 6a and b).

To examine whether scRobust can extract drug tolerance information after removing batch effects, we pre-trained scRobust with the integrated dataset after batch effect removal. We found that cell embeddings of scRobust were clustered mainly by treatment days, showing that scRobust can extract drug tolerance

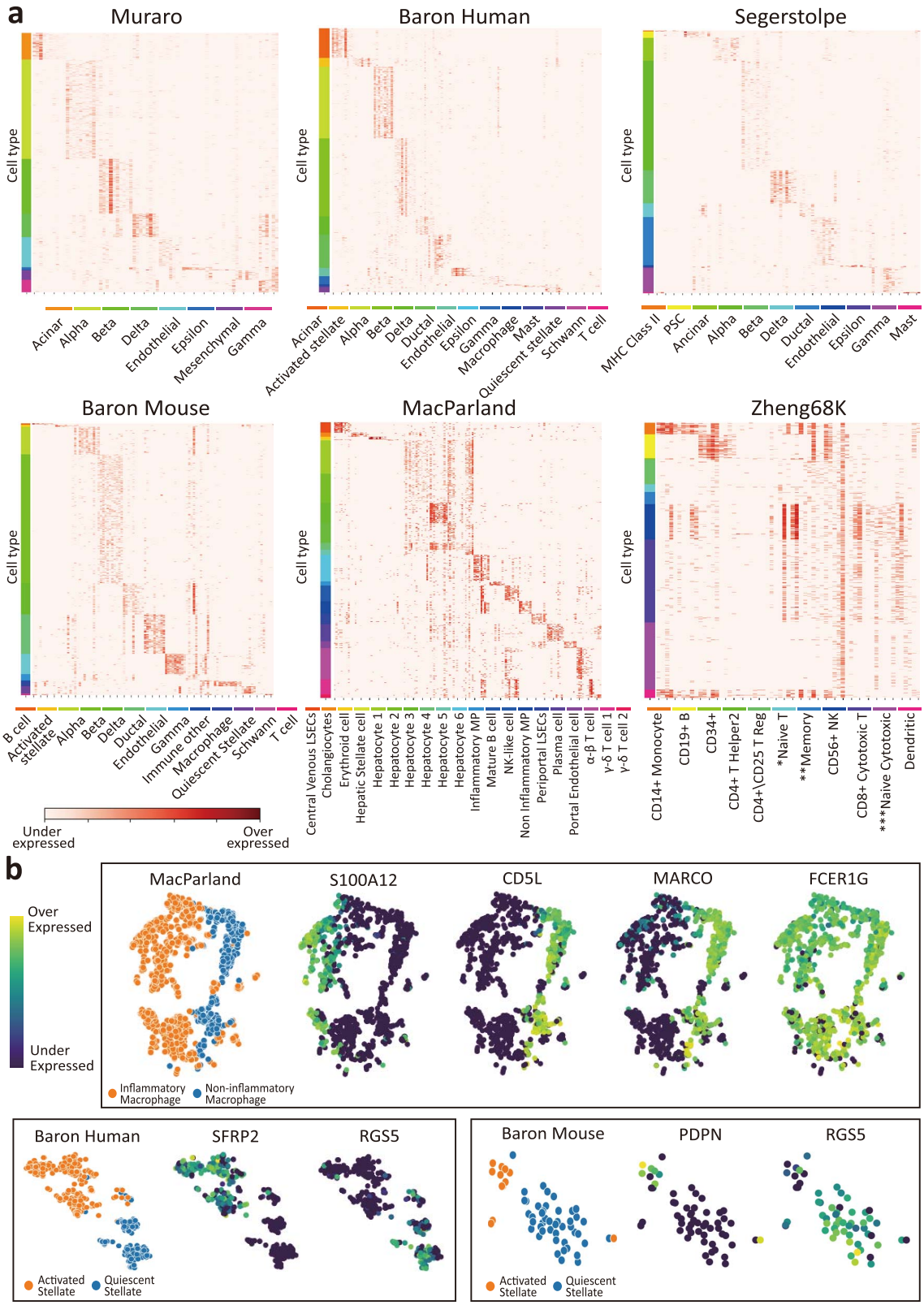


Figure 5. **Marker gene discovery by scRobust. (a)** Heatmaps showing the expression of the top genes with high attention scores identified by scRobust, where the x-axis presents genes (Tables S21–29) and cells clustered by cell type. *, **, and *** denote CD4+/CD45RA+/CD25-, CD4+/CD45RO+, and CD8+/CD45RA+ cell types, respectively. This visualization aids in understanding how these genes are differentially expressed across various cell types. **(b)** t-SNE plots of target cell embeddings for scRobust, colored based on the expression of marker genes. These plots illustrate the distinction of cells sharing the same parent cell type. In the MacParland dataset, for instance, non-inflammatory and inflammatory macrophages are differentiated by the S100A12, CD5L, and MARCO genes. However, FCER1G provides information pertaining to the macrophage type. Similarly, in the Baron Human and Mouse datasets, the marker genes effectively separate activated from quiescent stellate cells.

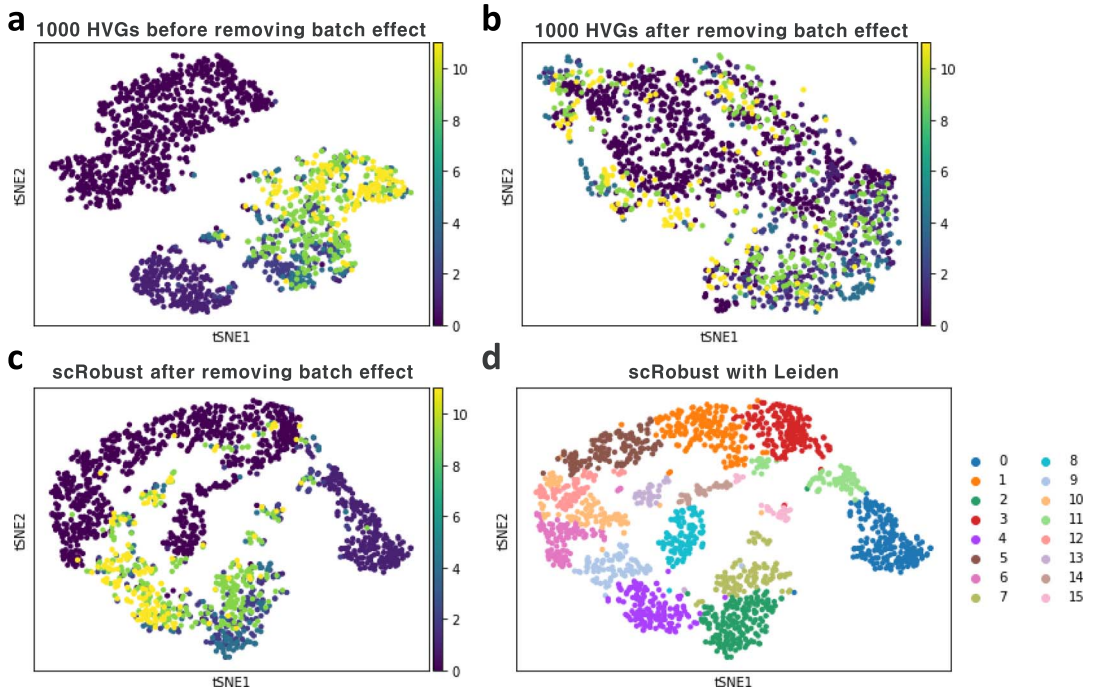


Figure 6. **t-SNE plots of integrated drug treatment datasets for days 0, 1, 2, 4, 9, and 11.** t-SNE plots of 1000 HVGs for the integrated six drug treatment datasets **(a)** before and **(b)** after removing the batch effect, colored by the days after treatment. t-SNE plots of cell embeddings of scRobust, colored by **(c)** the days after treatment and **(d)** cluster numbers identified by Leiden algorithm.

information (Fig. 6c). Furthermore, using the Leiden algorithm [32], we clustered the cell embeddings, and named clusters with Day 0 > 70% of cells as sensitive, Day 1 > 70% as early DTPs, Days 2 and 4 > 70% as DTPs, and Days 9 and 11 > 70% as DTEPs (Fig. 6d and Table S30). We used *t*-tests to identify marker genes among the stages, selecting the ten genes with the lowest *p*-values (Fig. S11). Comparing sensitive and resistant cells (early DTP, DTP, and DTEP) in our clusters (Table S31), we found that previously known drug resistance genes such as HSPB1, CYP1B1, SLC7A5, TAGLN, ENO1, ID3, MTRNR2L12, FAM134B, S100A6, MALAT1, ALDH3A1, and FTH1P3 were included, of which HSPB1, TAGLN, MTRNR2L12, ALDH3A1, and FTH1P3 were specific to lung cancer or EGFR inhibitors [37–49]. In addition, when using cell embeddings of scRobust, ALDH3A1, a DTP-associated gene [50], was detected as a marker gene between early DTPs and DTPs.

For comparison, we identified differentially expressed genes from drug-tolerance stages according to treatment days: all cells of Day 0 as sensitive, cells of day 1 as early DTPs, cells of days 2 and 4 as DTPs, and cells of days 9 and 11 as DTEPs. In this approach, although five previously known drug-resistance genes (CYP1B1, FSTL1 [51], SLC7A5, CCDC80 [52], and SERPINE1 [53]) were detected (Table S32), they were not specific to lung cancer or EGFR inhibitors. In summary, scRobust effectively identified marker genes distinguishing between sensitive and resistant cells and various stages of drug tolerance.

## scRobust in the scATAC-seq dataset

To verify whether our SSL strategy can be applied to other data types, we pre-trained scRobust on the PBMCs scATAC-seq dataset produced using the 10x Genomics Chromium system [3], which provided both PBMCs scATAC-seq and scRNA-seq data for the same cells. To assess the impact of pre-training, we compared t-SNE plots of cell embedding vectors generated by scRobust with those generated by PCA, using various numbers of input features (i.e. 200, 500, and 1,000 selected features). As shown in Fig. S12, all cell embedding vectors generated by scRobust using 200, 500, and 1000 features were clearly distinguished by cell types, whereas some cell embeddings from PCA using a small number of features were mixed. Specifically, CD16 Mono, CD14 Mono, and cDC cells were well-separated in scRobust across all cases, but these cell types were mixed in PCA when using 200 features. Additionally, we conducted the cell type annotation test using scATAC-seq and compared the results with those obtained using scRNA-seq (Fig. S13), where scRobust and CIForm used 1,000 and 2,000 features, respectively. scRobust using scATAC-seq achieved reasonable F1 and accuracy scores (F1: 0.789 and Acc: 0.883) compared with scRobust and CIForm using scRNA-seq (scRobust: F1: 0.865 and Acc: 0.909; CIForm: F1: 0.835 and Acc: 0.890), whereas CIForm using scATAC-seq showed inferior performance (F1: 0.657 and Acc: 0.800). These results demonstrate that scRobust can still extract global information from a small number of features in scATAC-seq data and has the potential to interact effectively with various omics data types.

## Discussion

Our SSL strategy facilitates the learning of all genes and their interrelations, rather than just a selected subset. This methodology enables the encoder to access and integrate global information. Consequently, scRobust demonstrated superior performance in cell-type annotation over benchmark methods, especially in the annotation of rare cell types and additional dropout scenarios (Fig. 2). scRobust showed remarkable performance under various conditions (Fig. 3). Comparing cell embeddings between methods revealed that scRobust effectively captured both cell-type and sample-specific information, such as glucose control parameters in diabetes. In marker gene discovery, scRobust effectively identified marker genes across various cell types (Fig. 5).

scRobust employed highly unique genes unlike benchmark methods using HVGs. This increases the chance of using marker genes, which tend to be over- or exclusively expressed in the target cells. Consequently, scRobust incorporated many T-cell marker genes in the Baron mouse dataset and far outperformed the benchmark methods.

# Methods

## Architecture of scRobust

The encoder of scRobust, modeled after the Transformer [54], adapts its number of layers based on the dataset size. For larger datasets, such as Zheng 68K [3], Zheng sorted [3], and TM [22], with cell counts of 68,579, 20,000, and 54,865, respectively, a two-layer encoder is employed. However, the encoder for the other datasets with relatively small cells comprises a single layer. The number of attention heads is consistently set to eight across all datasets. Unlike the original BERT model where words are in order, scRobust omits segment and position embeddings as gene order is not influential; it solely utilizes token (gene) embedding vectors.

Gene embedding vectors in scRobust are initialized as trainable and random, differing from gene embeddings such as Gene2vec [55] used in other BERT-based cell type annotation models, such as scBERT [5]. The gene vocabulary in scRobust varies depending on the dataset and includes five special tokens: "PAD," "SEP," "UNKNOWN," "CLS," and "MASK." Input tokens comprise $m$ genes along with the CLS token, which serves as a summary vector. The output from the "CLS" token is treated as a cell-embedding vector, utilized in contrastive learning, gene expression prediction, and cell type annotation. In our experiments, the gene-embedding dimension was set to 512.

The gene-embedding vectors and gene-expression values for all samples are represented as $(\mathcal{E}, \mathcal{G})$, where $\mathcal{E} = \mathbf{E}_1, \mathbf{E}_2, ..., \mathbf{E}_{|\mathcal{E}|}$ and $\mathcal{G} = \mathbf{GE}_1, \mathbf{GE}_2, ..., \mathbf{GE}_{|\mathcal{G}|}$. Here, $\mathbf{E}_j \in \mathbb{R}^d$ and $\mathbf{GE}_i = [\mathbf{ge}_1^i, \mathbf{ge}_2^i, ..., \mathbf{ge}_{|\mathbf{E}|}^i] \in \mathbb{R}^{|\mathbf{E}|}$ denote the embedding vector for gene $j$, where $|\mathbf{E}| = 1$, and the gene expression vector of sample $i$, where $\mathbf{ge}_j^i \in \mathbb{R}$ represents the expression value of gene $j$ in sample $i$, respectively.

To incorporate gene-expression values into gene-embedding vectors, we normalized the gene-embedding vectors to unit vectors and scaled them with the corresponding gene-expression values, similar to the approach for GEN [56]. For example, the embedding vector for gene $j$ in sample $i$ is scaled as $\mathbf{ge}_j^i \times \mathbf{E}_j$. Thus, the input and output tokens for sample $i$ can be represented as follows:

$$X_i = [\mathbf{E}_{[CLS]}, \mathbf{ge}_1^i \times \mathbf{E}_1, \mathbf{ge}_2^i \times \mathbf{E}_2, ..., \mathbf{ge}_m^i \times \mathbf{E}_m] \quad (1)$$

$$[h_{[CLS]}^i, h_1, ...h_m] = \mathbf{EnC}(X_i) \quad (2)$$

where $h \in \mathbb{R}^d$, m is the number of input genes, EnC denotes the Transformer encoder, and $h_{[CLS]}^i$ is a summary vector representing cell embedding.

Unlike conventional methods using gene expression data, scRobust does not require the same genes as inputs for all cells, thanks to gene embeddings being represented by vectors, not fixed input vector indices. For instance, if scRobust generates two embedding vectors from two random gene sets, these vectors would represent the cell embeddings for the target cell.

## Contrastive learning

In a mini-batch containing $N$ cells, scRobust creates two distinct cell-embedding vectors for each cell by utilizing two subsets of $m$ randomly selected genes, which generates a total of $2N$ cell-embedding vectors. Within this set, the vector pair originating from the same cell is considered the "positive pair," whereas the remaining $2(N-1)$ vectors, derived from other cells, are treated as "negative samples." The formulation of two different cell-embedding vectors for a given sample $i$ can be described as follows:

$$X_{i_1} = [\mathbf{E}_{[CLS]}, \mathbf{ge}_{r_1}^i \times \mathbf{E}_{r_1}, \mathbf{ge}_{r_2}^i \times \mathbf{E}_{r_2}, ..., \mathbf{ge}_{r_m}^i \times \mathbf{E}_{r_m}] \quad (3)$$

$$X_{i_2} = [\mathbf{E}_{[CLS]}, \mathbf{ge}_{\tilde{r}_1}^i \times \mathbf{E}_{\tilde{r}_1}, \mathbf{ge}_{\tilde{r}_2}^i \times \mathbf{E}_{\tilde{r}_2}, ..., \mathbf{ge}_{\tilde{r}_m}^i \times \mathbf{E}_{\tilde{r}_m}] \quad (4)$$

$$[h_{[CLS]}^{i_1}, h_{r_1}, ...h_{r_m}] = \mathbf{EnC}(X_{i_1}), [h_{[CLS]}^{i_2}, h_{\tilde{r}_1}, ...h_{\tilde{r}_m}] = \mathbf{EnC}(X_{i_2}), \quad (5)$$

where $r \in \mathbb{R}^m$ and $\tilde{r} \in \mathbb{R}^m$ represent vectors of random indices, $m$ denotes the number of input genes, and EnC denotes the Transformer encoder.

We used a fully connected network (FCN) to project the cell-embedding vector $h_{[CLS]}$ into the cell embedding space. The contrastive loss function was identical to that used in simCLR [11], facilitating the effective differentiation of cells in the embedding space.

$$z_{i_1} = f\left(h_{[CLS]}^{i_1}\right) \quad (6)$$

$$L_{cl_{i_1,i_2}} = -\log \frac{\exp(\mathrm{sim}(z_{i_1}, z_{i_2})/\tau)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i_2]} \exp(\mathrm{sim}(z_{i_1}, z_k)/\tau)} \quad (7)$$

Here, $f$ represents the FCN with two layers, $\mathbf{1}$ is an indicator function that evaluates to 1 if $k \neq i_1$, and $\tau$ is a temperature parameter, which was set to 0.07 in our experiment.

In contrastive learning, we used approximately 10% of the average number of genes without dropout as input genes (e.g. 50, 100, or 200 genes); thus a relatively small number of input genes were used for pre-training. Notably, this approach not only conserves computational resources but also enhances the quality of the pre-trained model.

## Gene expression prediction task

The cell-embedding vector, which successfully captures global information, should encompass the complete gene expression profile of the cell. To encode the cell-embedding vectors, we trained the encoder to predict gene expression values from the dot product between the projected cell-embedding vector and corresponding gene-embedding vectors. The accuracy of these predictions was assessed using mean squared error (MSE) as the loss function.

$$\tilde{h}_{i_1} = g\left(h_{[CLS]}^i\right) \quad (8)$$

$$[\hat{\mathbf{ge}}_{r_1}^i, \hat{\mathbf{ge}}_{r_2}^i, ..., \hat{\mathbf{ge}}_{r_m}^i] = \tilde{h}_{i_1}^T [\mathbf{E}_{r_1}, \mathbf{E}_{r_2}, ..., \cdot \mathbf{E}_{r_m}] \quad (9)$$

$$L_{ge} = \frac{1}{m} \sum_{k=1}^{m} (\mathbf{ge}_{r_k}^i - \hat{\mathbf{ge}}_{r_k}^i)^2 \quad (10)$$

Here, $g$ is the FCN with two layers and $r \in \mathbb{R}^m$ is a random-index vector, which differs from that used to generate the cell-embedding vector.

## Objective function of scRobust

In the pre-training stage, contrastive learning and gene expression prediction tasks progress simultaneously, and the combined self-supervised learning loss can be described as follows:

$$L_{ssl} = L_{cl} + L_{ge}. \quad (11)$$

For each dataset, we constructed and trained separate encoders due to the variability in gene numbers and sequencing platforms. This approach necessitates distinct gene vocabularies and pre-trained models for each dataset. After the pre-training stage, we fine-tuned the encoders for cell type annotation.

$$X_i = [\mathbf{E}_{[CLS]}, \mathbf{ge}^i_{u_1} \times \mathbf{E}_{u_1}, \mathbf{ge}^i_{u_2} \times \mathbf{E}_{u_2}, ..., \mathbf{ge}^i_{u_n} \times \mathbf{E}_{u_n}] \quad (12)$$

$$[h^i_{[CLS]}, h_{u_1}, ...h_{u_n}] = \mathbf{EnC}(X_i) \quad (13)$$

$$\hat{t} = c(h^i_{[CLS]}) \quad (14)$$

$$L = \mathbf{CrossEntropy}(t, \hat{t}) \quad (15)$$

Here, $u \in \mathbb{R}^n$ represents an index vector for highly unique genes, EnC denotes the Transformer encoder, $c$ denotes a classifier based on the FCN with two layers, and $t$ and $\hat{t}$ correspond to the true and predicted cell types, respectively. By utilizing a larger number of genes in the downstream process than in the pre-training stage, the representational power of the cell embedding is maximized. Accordingly, we used 800 highly unique genes in our experiments.

## Unique gene selection

As scRobust can use cell-specific input genes, rarely expressed genes (potential marker genes) were prioritized as input. First, for each gene of the scRNA-seq dataset, we calculated the proportion of zero values (dropout) across samples. Second, for each sample, we sorted genes in ascending proportions of zero values, prioritizing rarely expressed genes. Third, for each sample, input genes were selected based on their order, prioritizing the most unique $N$ genes in each sample. The input genes differed for each sample because each cell had different non-zero read count genes. The use of highly unique genes can maximize the unique features of each cell. Although scRobust does not use the same genes for all cells, it can extract common features among cells because pre-training of the encoder considers various subsets of whole genes. Therefore, scRobust uses individual gene sets consisting of highly unique genes in downstream tasks.

## Pathway embedding vector

As pathways are defined by specific gene sets, scRobust can generate pathway-related embedding vectors based on these gene sets. In this context, we define two types of vectors: pathway-informed cell embedding vectors and pathway embedding vectors. A pathway-informed cell embedding vector is a type of cell embedding vector that utilizes the pathway gene set and the gene expression values of an input cell. Similarly, a pathway embedding vector also relies on the genes within the pathway, but these gene embedding vectors are not scaled by gene expression values. These embedding vectors can be described as follows:

$$X^i_{pw} = [\mathbf{E}_{[CLS]}, \mathbf{ge}^i_{pw_1} \times \mathbf{E}_{pw_1}, \mathbf{ge}^i_{pw_2} \times \mathbf{E}_{pw_2}, ..., \mathbf{ge}^i_{pw_m} \times \mathbf{E}_{pw_m}] \quad (16)$$

$$X_{pw} = [\mathbf{E}_{[CLS]}, \mathbf{E}_{pw_1}, \mathbf{E}_{pw_2}, ..., \mathbf{E}_{pw_m}] \quad (17)$$

$$[h^{pw,i}_{[CLS]}, h^i_{pw_1}, ...h^i_{pw_m}] = \mathbf{EnC}(X^i_{pw}) \quad (18)$$

$$[h^{pw}_{[CLS]}, h_{pw_1}, ...h_{pw_m}] = \mathbf{EnC}(X_{pw}), \quad (19)$$

where $pw_k$ is gene $k$ of the pathway gene set, $i$ represents cell $i$, and $h^{pw,i}_{[CLS]}$ and $h^{pw}_{[CLS]}$ are pathway-informed cell embedding and pathway embedding vectors, respectively.

Since pathway embedding vectors are defined in the same latent space as cell embedding vectors, we can assess how close a pathway-informed cell embedding vector is to a target pathway vector. To evaluate the closeness between two vectors, both the direction and magnitude of the vectors must be considered, as not only the direction but also the length of the vectors are important in the cell embedding space. Therefore, we define the following formula to evaluate how close vector **a** is to vector **b**:

$$\text{closeness}(a)_b = \frac{a \cdot b}{b \cdot b} = cos(\theta)\frac{|a|}{|b|}, \quad (20)$$

where $\theta$ is the angle between vectors a and b.

To decide the cell type of a given cell, we calculate the closeness scores between the input cell and candidate cell type pathway vectors. Then, we annotate the cell with the cell type having the highest closeness among candidate pathways.

**Key Points**

- Through the novel self-supervised learning strategy, scRobust tackles the sparsity inherent in single-cell RNA-seq data by learning all genes' information.
- scRobust achieves state-of-the-art performance on eight out of nine benchmark datasets on cell-type annotation tasks and in a variety of scenarios.
- scRobust can predict cell types using pathway vectors without fine-tuning and clustering.
- scRobust can generate high-quality cell embeddings including not only cell-specific but also sample-specific information.
- scRobust can detect the marker genes for cell types and various drug tolerance stages.

## Acknowledgements

This work was supported by an Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2019-0-00567, Development of Intelligent SW Systems for Uncovering Genetic Variation and Developing Personalized Medicine for Cancer Patients with Unknown Molecular Genetic Mechanisms and No.2019-0-01842, Artificial Intelligence Graduate School Program (GIST)).

## Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

## Competing interests statement

There are no competing interests.

## Data availability

The PBMC datasets, Zheng 68K and Zheng sorted, were downloaded from https://support.10xgenomics.com/single-cell-gene-expression/datasets. The pancreas datasets were downloaded from different sources: Baron from GEO with accession number GSE84133, Muraro from GSE85241, Segerstolpe from ArrayExpress with accession number E-MTAB-5061, and Xin from GSE81608. The MacParland liver dataset was sourced from GSE115469. The TM mouse cell dataset was downloaded from GSE109774. All

these datasets are publicly accessible. The processed datasets utilized in our study are available on Zenodo (https://zenodo.org/records/10602754).

## Code availability

The source code for scRobust and pre-trained weights are available on GitHub https://github.com/DMCB-GIST/scRobust.

## Author contributions

H.L. initiated the study and contributed to the study's concept and design. S.P. designed and implemented the proposed algorithms. H.L. and S.P. analyzed and interpreted the results. H.L. and S.P. wrote the manuscript. H.L. took part in the study supervision and coordination. All authors reviewed the manuscript.

## References


1. Satija R, Farrell JA, Gennert D. *et al.* Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 2015;**33**:495–502. https://doi.org/10.1038/nbt.3192.
2. Wang J, Wang Z, Yuan S. *et al.* A clustering method for single-cell rna-seq data based on automatic weighting penalty and low-rank representation. *IEEE/ACM Trans Comput Biol Bioinform* 2024;**21**:360–71. https://doi.org/10.1109/TCBB.2024.3362472.
3. Zheng GXY, Terry JM, Belgrader P. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 2017;**8**:14049. https://doi.org/10.1038/ncomms14049.
4. Picelli S, Björklund ÅK, Faridani OR. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* 2013;**10**:1096–8. https://doi.org/10.1038/nmeth.2639.
5. Yang F, Wang W, Wang F. *et al.* Scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nat Mach Intell* 2022;**4**:852–66. https://doi.org/10.1038/s42256-022-00534-z.
6. Chen J, Hao X, Tao W. *et al.* Transformer for one stop interpretable cell type annotation. *Nat Commun* 2023;**14**:223. https://doi.org/10.1038/s41467-023-35923-4.
7. Jing X, Zhang A, Liu F. *et al.* Ciform as a transformer-based model for cell-type annotation of large-scale single-cell rna-seq data. *Brief Bioinform* 2023;**24**:bbad195.
8. Cui H, Wang C, Maan H. *et al.* Scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nat Methods* 2024;**21**:1470–1480.
9. Dosovitskiy A, Beyer L, Kolesnikov A. *et al.* An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. 2020.
10. Hao M, Gong J, Zeng X. *et al.* Large-scale foundation model on single-cell transcriptomics. *Nat Methods* 2024;**21**:1481–1491.
11. Chen T, Kornblith S, Norouzi M. *et al.* A simple framework for contrastive learning of visual representations. In: Iii HD, Singh A. (eds). *International Conference on Machine Learning*. Proceedings of Machine Learning Research, Virtual, pp. 1597–607. 2020.
12. Gao T, Yao X, Chen D. Simcse: simple contrastive learning of sentence embeddings. arXiv preprint arXiv:2104.08821. 2021.
13. Qian R, Meng T, Gong B. *et al.* Spatiotemporal contrastive video representation learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. IEEE Computer Society, Washington, DC, 6964–74, 2021.
14. Yang M, Yang Y, Xie C. *et al.* Contrastive learning enables rapid mapping to multimodal single-cell atlas of multimillion scale. *Nat Mach Intell* 2022;**4**:696–709. https://doi.org/10.1038/s42256-022-00518-z.
15. Han W, Cheng Y, Chen J. *et al.* Self-supervised contrastive learning for integrative single cell rna-seq data analysis. *Brief Bioinform* 2022;**23**. https://doi.org/10.1093/bib/bbac377.
16. Liu J, Zeng W, Kan S. *et al.* Cake: a flexible self-supervised framework for enhancing cell visualization, clustering and rare cell identification. *Brief Bioinform* 2024;**25**:bbad475.
17. Malkov YA, Yashunin DA. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Trans Pattern Anal Mach Intell* 2018;**42**:824–36. https://doi.org/10.1109/TPAMI.2018.2889473.
18. Baron M, Veres A, Wolock SL. *et al.* A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell Syst* 2016;**3**:346–360.e4. https://doi.org/10.1016/j.cels.2016.08.011.
19. Muraro MJ, Dharmadhikari G, Grün D. *et al.* A single-cell transcriptome atlas of the human pancreas. *Cell Syst* 2016;**3**:385–394.e3. https://doi.org/10.1016/j.cels.2016.09.002.
20. Segerstolpe Å, Palasantza A, Eliasson P. *et al.* Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab* 2016;**24**:593–607. https://doi.org/10.1016/j.cmet.2016.08.020.
21. Xin Y, Kim J, Okamoto H. *et al.* Rna sequencing of single human islet cells reveals type 2 diabetes genes. *Cell Metab* 2016;**24**:608–15. https://doi.org/10.1016/j.cmet.2016.08.018.
22. Schaum N, Karkanias J, Neff NF. *et al.* Single-cell transcriptomics of 20 mouse organs creates a tabula muris: the tabula muris consortium. *Nature* 2018;**562**:367.
23. MacParland SA, Liu JC, Ma X-Z. *et al.* Single cell rna sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat Commun* 2018;**9**:4383. https://doi.org/10.1038/s41467-018-06318-7.
24. Okonechnikov K, Joshi P, Sepp M. *et al.* Mapping pediatric brain tumors to their origins in the developing cerebellum. *Neuro Oncol* 2023;**25**:1895–909. https://doi.org/10.1093/neuonc/noad124.
25. Schirmer L, Velmeshev D, Holmqvist S. *et al.* Neuronal vulnerability and multilineage diversity in multiple sclerosis. *Nature* 2019;**573**:75–82. https://doi.org/10.1038/s41586-019-1404-z.
26. Hodge RD, Bakken TE, Miller JA. *et al.* Conserved cell types with divergent features in human versus mouse cortex. *Nature* 2019;**573**:61–8. https://doi.org/10.1038/s41586-019-1506-7.
27. Lake BB, Menon R, Winfree S. *et al.* An atlas of healthy and injured cell states and niches in the human kidney. *Nature* 2023;**619**:585–94. https://doi.org/10.1038/s41586-023-05769-3.
28. Knight-Schrijver VR, Davaapil H, Bayraktar S. *et al.* A single-cell comparison of adult and fetal human epicardium defines the age-associated changes in epicardial activity. *Nat Cardiovasc Res* 2022;**1**:1215–29. https://doi.org/10.1038/s44161-022-00183-w.
29. Domínguez, Conde C, Xu C, Jarvis LB. *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* 2022;**376**:eabl5197. https://doi.org/10.1126/science.abl5197.
30. Lake BB, Ai R, Kaeser GE. *et al.* Neuronal subtypes and diversity revealed by single-nucleus rna sequencing of the human brain. *Science* 2016;**352**:1586–90. https://doi.org/10.1126/science.aaf1204.
31. Subramanian A, Tamayo P, Mootha VK. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* 2005;**102**:15545–50.
32. Traag VA, Waltman L, Van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* 2019;**9**:5233.

33. Lira-Junior R, Holmström SB, Clark R. *et al.* S100a12 expression is modulated during monocyte differentiation and reflects periodontitis severity. *Front Immunol* 2020;**11**:86. https://doi.org/10.3389/fimmu.2020.00086.
34. La Fleur L, Botling J, He F. *et al.* Targeting Marco and il37r on immunosuppressive macrophages in lung cancer blocks regulatory t cells and supports cytotoxic lymphocyte function. *Cancer Res* 2021;**81**:956–67. https://doi.org/10.1158/0008-5472.CAN-20-1885.
35. Aissa AF, Islam ABMMK, Ariss MM. *et al.* Single-cell transcriptional changes associated with drug tolerance and response to combination therapies in cancer. *Nat Commun* 2021;**12**:1628. https://doi.org/10.1038/s41467-021-21884-z.
36. Tran HTN, Ang KS, Chevrier M. *et al.* A benchmark of batch-effect correction methods for single-cell rna sequencing data. *Genome Biol* 2020;**21**:1–32. https://doi.org/10.1186/s13059-019-1850-9.
37. Lelj-Garolla B, Kumano M, Beraldi E. *et al.* Hsp27 inhibition with ogx-427 sensitizes non–small cell lung cancer cells to erlotinib and chemotherapy. *Mol Cancer Ther* 2015;**14**:1107–16. https://doi.org/10.1158/1535-7163.MCT-14-0866.
38. Chen P, Wang S, Cao C. *et al.* $\alpha$-Naphthoflavone-derived cytochrome P450 (CYP)1B1 degraders specific for sensitizing CYP1B1-mediated drug resistance to prostate cancer DU145: Structure activity relationship. *Bioorg Chem* 2021;**116**:105295. https://doi.org/10.1016/j.bioorg.2021.105295.
39. Yoo H-C, Han J-M. Amino acid metabolism in cancer drug resistance. *Cells* 2022;**11**:140. https://doi.org/10.3390/cells11010140.
40. Kim I-G, Lee J-H, Kim S-Y. *et al.* Hypoxia-inducible transgelin 2 selects epithelial-to-mesenchymal transition and $\gamma$-radiation-resistant subtypes by focal adhesion kinase-associated insulin-like growth factor 1 receptor activation in non-small-cell lung cancer cells. *Cancer Sci* 2018;**109**:3519–31. https://doi.org/10.1111/cas.13791.
41. Jinrong G, Zhong K, Wang L. *et al.* Eno1 contributes to 5-fluorouracil resistance in colorectal cancer cells via emt pathway. *Front Oncol* 2022;**12**:1013035. https://doi.org/10.3389/fonc.2022.1013035.
42. Larribère L, Novak D, Huizi W. *et al.* New role of id3 in melanoma adaptive drug-resistance. *Oncotarget* 2017;**8**:110166–75. https://doi.org/10.18632/oncotarget.22698.
43. Lin P, Cheng W, Qi X. *et al.* Bioinformatics and experimental validation for identifying biomarkers associated with amg510 (sotorasib) resistance in krasg12c-mutated lung adenocarcinoma. *Int J Mol Sci* 2024;**25**:1555. https://doi.org/10.3390/ijms25031555.
44. Chipurupalli S, Desiderio V, Robinson N. Analysis of er-phagy in cancer drug resistance. In: Baiocchi M. (eds). *Cancer Drug Resistance: Methods and Protocols*. Springer, Humana, New York, NY, 2022, 211–220, https://doi.org/10.1007/978-1-0716-2513-2_16.
45. Luo T, Liu Q, Tan A. *et al.* Mesenchymal stem cell-secreted exosome promotes chemoresistance in breast cancer via enhancing mir-21-5p-mediated s100a6 expression. *Mol Ther Oncolytics* 2020;**19**:283–93. https://doi.org/10.1016/j.omto.2020.10.008.
46. Hou J, Zhang G, Wang X. *et al.* Functions and mechanisms of lncrna malat1 in cancer chemotherapy resistance. *Biomark Res* 2023;**11**:23. https://doi.org/10.1186/s40364-023-00467-8.
47. Kumar S, Mishra S. Malat1 as master regulator of biomarkers predictive of pan-cancer multi-drug resistance in the context of recalcitrant nras signaling pathway identified using systems-oriented approach. *Sci Rep* 2022;**12**:7540. https://doi.org/10.1038/s41598-022-11214-8.
48. Rebollido-Rios R, Venton G, Sánchez-Redondo S. *et al.* Dual disruption of aldehyde dehydrogenases 1 and 3 promotes functional changes in the glutathione redox system and enhances chemosensitivity in nonsmall cell lung cancer. *Oncogene* 2020;**39**:2756–71. https://doi.org/10.1038/s41388-020-1184-9.
49. Zheng G, Chen W, Li W. *et al.* E2f1-induced ferritin heavy chain 1 pseudogene 3 (fth1p3) accelerates non-small cell lung cancer gefitinib resistance. *Biochem Biophys Res Commun* 2020;**530**:624–31. https://doi.org/10.1016/j.bbrc.2020.07.044.
50. Chen M, Mainardi S, Lieftink C. *et al.* Targeting of vulnerabilities of drug-tolerant persisters identified through functional genetics delays tumor relapse. *Cell Rep Med* 2024;**5**:101471. https://doi.org/10.1016/j.xcrm.2024.101471.
51. Nie E, Miao F, Jin X. *et al.* Fstl1/dip2a/mgmt signaling pathway plays important roles in temozolomide resistance in glioblastoma. *Oncogene* 2019;**38**:2706–21. https://doi.org/10.1038/s41388-018-0596-2.
52. Wang W-D, Guo-Yan W, Bai K-H. *et al.* A prognostic stemness biomarker ccdc80 reveals acquired drug resistance and immune infiltration in colorectal cancer. *Clin Transl Med* 2020;**10**:e225. https://doi.org/10.1002/ctm2.225.
53. Zhang Q, Lei L, Jing D. Knockdown of serpine1 reverses resistance of triple-negative breast cancer to paclitaxel via suppression of vegfa. *Oncol Rep* 2020;**44**:1875–84. https://doi.org/10.3892/or.2020.7770.
54. Vaswani A, Shazeer N, Parmar N. *et al.* Attention is all you need. In: Guyon I. *et al.* (eds). *Adv Neural Inf Process Syst* Curran Associates, Inc., Red Hook, NY, 2017;**30**:5998–6008.
55. Jingcheng D, Jia P, Dai Y. *et al.* Gene2vec: distributed representation of genes based on co-expression. *BMC Genom* 2019;**20**:7–15.
56. Park S, Lee H. Molecular data representation based on gene embeddings for cancer drug response prediction. *Sci Rep* 2023;**13**:21898. https://doi.org/10.1038/s41598-023-49003-6.