Front. Comput. Sci., 2023, 17(3): 173902
https://doi.org/10.1007/s11704-022-2011-y

RESEARCH ARTICLE

# AE-TPGG: a novel autoencoder-based approach for single-cell RNA-seq data imputation and dimensionality reduction

**Shuchang ZHAO**[1,2], **Li ZHANG**[1,3], **Xuejun LIU** (✉)[1,2]

1 MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
2 Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 210023, China
3 College of Computer Science and Technology, Nanjing Forestry University, Nanjing 210037, China



**Abstract** Single-cell RNA sequencing (scRNA-seq) technology has become an effective tool for high-throughout transcriptomic study, which circumvents the averaging artifacts corresponding to bulk RNA-seq technology, yielding new perspectives on the cellular diversity of potential superficially homogeneous populations. Although various sequencing techniques have decreased the amplification bias and improved capture efficiency caused by the low amount of starting material, the technical noise and biological variation are inevitably introduced into experimental process, resulting in high dropout events, which greatly hinder the downstream analysis. Considering the bimodal expression pattern and the right-skewed characteristic existed in normalized scRNA-seq data, we propose a customized autoencoder based on a two-part-generalized-gamma distribution (AE-TPGG) for scRNA-seq data analysis, which takes mixed discrete-continuous random variables of scRNA-seq data into account using a two-part model and utilizes the generalized gamma (GG) distribution, for fitting the positive and right-skewed continuous data. The adopted autoencoder enables AE-TPGG to captures the inherent relationship between genes. In addition to the ability of achieving low-dimensional representation, the AE-TPGG model also provides a denoised imputation according to statistical characteristic of gene expression. Results on real datasets demonstrate that our proposed model is competitive to current imputation methods and ameliorates a diverse set of typical scRNA-seq data analyses.

**Keywords** scRNA-seq, autoencoder, TPGG, data imputation, dimensionality reduction

## 1 Introduction

Single-cell RNA sequencing (scRNA-seq) has recently rung in rapid innovation with the emergence of various technologies, leading to high throughput and facilitating dissection of heterogeneity in cell populations [1]. In terms of technical process, scRNA-seq involves reverse transcription and preparation of a cDNA library followed by high-throughput DNA sequencing, which is fundamentally similar to traditional bulk RNA-seq methods. The most significant difference is that it allows the dissection of gene expression at single-cell resolution instead of one single library from the RNA pool of many cells. One study that adopts scRNA-seq to study gene expression in human colorectal tumours and matched normal mucosa identifies *TGFB*1 to be the most upregulated differentially expressed regulatory gene in cancer-associated fibroblasts (CAFs) [2]. Under the COVID-19 pandemic sweeping the world, high-throughput single-cell RNA and VDJ sequencing of antigen-enriched B cells from 60 convalescent patients are utilized to rapidly recognize the SARS-CoV-2-neutralizing antibodies [3]. However, duo to the low amount of starting mRNA content, scRNA-seq technologies suffer from many sources of significant technical noise, the most prominent of which is the dropout events caused by inefficient mRNA capture [4−6]. Therefore, the data imputation is conducive to reveal the expression pattern of the original biological signal submerged by the noise.

The imputation methods of scRNA-seq data have advanced accompanied by the emergence of issues [7−10]. However, most of these methods are based on the correlation structure of single-cell expression profile to impute missing values by utilizing information on similarities between cells and/or genes, which may fail to account for the complexity and nonlinearity in the data. Deep-learning methods learn from multiple levels of representation by composing simple but non-linear modules that each transforms the representation at one level into a representation at a higher abstract level [11]. Autoencoders, as a kind of artificial neural networks, are an unsupervised learning paradigm, which can learn the effective low dimensional representation of data, generally by minimizing the reconstructed error between the original input and decoded output [12]. The universal approximation theorem guarantees that the forward neural network with at least one hidden layer and enough neurons can approximate any function with any accuracy [13], which means that the

Received January 7, 2022; accepted April 15, 2022

E-mail: xuejun.liu@nuaa.edu.cn

autoencoders can realize almost perfect reconstruction of the input. In addition, autoencoders can extract meaningful features of data by applying different regularities to remove the random noise or redundant information inside the data. As deep networks produce better compression than shallow or linear networks [14], the autoencoders have been used in the field of life sciences to enhance the biological information mining [15,16]. Therefore, developing an autoencoder to account for the characteristic of the scRNA-seq data is the focus of this work.

As an important step of data preprocessing, normalization can effectively alleviate a series of biological and technical variations during data generation [17]. It is well known that compared with the traditional transcriptome sequencing technology, the preparation of single-cell transcriptome data library requires a higher level of polymerase chain reaction (PCR) amplification, and the bias is more serious with the increase of amplification times. This bias will result in a series of variations including those across cells, such as library size [18], or those within-cell, for example, guanine-cytosine (GC) content [19,20], gene sequence length, or unexpected variation introduced by batch effect. Thus, normalization is a critical step in the analysis pipeline to adjust for unexpected biological and technical effects that can mask the signal of interest [21]. There are many methods of data normalization for scRNA-seq data, such as Reads Per Million (RPM), which standardizes the total number of reads between cells [21]. In addition, Reads Per Kilobase Million (RPKM) [18] and Transcripts Per Million (TPM) [22] are also the common normalization strategies for gene expression analysis. After normalization, the original count data changes from discrete data to semi-continuous data. A large number of zero expression and positive expression present a characteristic bimodal expression pattern. In addition, we find through the data analysis in Section 1 that the continuous data has a typical right-skewed characteristic, which is still under insufficient exploration for current imputation methods.

In this work, we propose AE-TPGG, a novel autoencoder-based model combined with the Two-Part-Generalized-Gamma (TPGG) distribution accounting for semi-continuity of normalized scRNA-seq data. The motivation of the selection of TPGG derives from the two statistical features presented by the normalized data: the bimodal expression pattern and the right-skewed characteristic of positive expression. The bimodal expression pattern has two basic statistical features of the observed $x$: 1) $x_i \geqslant 0$, and 2) $x_i = 0$ is observed often enough that there are absorbing substantive and statistical properties for special consideration [23]. Owing to the mass point at zero, a single model for such data may be undesirable. The two-part model considers the mass of zeros by employing the Bernoulli distribution to fit for the probability of observing a positive-versus-zero outcome. For the right-skewed characteristic of positive expression, a flexible distribution in statistical literature, i.e., the generalized gamma (GG) distribution, can adaptively perform the model selection for two alternative right-skewed distributions of gamma and lognormal [24]. In the training of the autoencoder, we adopt the negative log-likelihood of TPGG rather than the most commonly used mean squared error (MSE) as the loss to inference gene-specific distribution parameters. Furthermore, the connectivity of the neural network naturally accounts for the inter-dependent relationships between genes that are not considered in the traditional maximum likelihood estimation.

Our contribution in this work is threefold: (1) We introduce a TPGG distribution for modeling the gene expression that accounts for the bimodal expression mode of the normalized scRNA-seq data and achieves adaptive model selection of gamma and lognormal distributions for the right-skewed distribution of positive expression. (2) The individualized autoencoder based on TPGG, namely AE-TPGG, utilizes loss of gene-specific distribution to substitute for conventional MSE loss. Not only is the low dimensional representation of cells obtained, but the parameters of distribution can be inferred, leading to data imputation according to first order origin moment of TPGG distribution. (3) Empirical analyses using real datasets demonstrate that AE-TPGG improves various analyses of scRNA-seq data. Specially, compared with popular methods of scRNA-seq data imputation, our method obtains competitive advantages in many metrics.

## 2 Related work

Considering the distribution characteristic of scRNA-seq data has always been the key to model gene expression. Different from the traditional bulk RNA-seq data, the gene expression level of single cell shows a bimodal expression pattern. Therefore, the traditional over-dispersed negative-binomial (NB) distribution that is commonly used in bulk-cell RNA-sequencing is unsuitable for modeling scRNA-seq data. Davide et. al. assumed that the count data follows zero-inflated negative binomial model that takes zero inflation, over-dispersion, and the discrete nature of the count data into account [25]. For the semi-continuity of the normalized scRNA-seq data, a two-part generalized regression model was developed to fit scRNA-seq data by Finak et. al. [5], which used logistic regression models for discrete variables and a Gaussian linear model for continuous variables. In addition to the bimodal expression mode of scRNA-seq data, we further observe that the positive expression presents a typical right-skewed characteristic (see detailed data analysis in Section 1), which is still under-explored for modeling gene expression.

In addition, due to high dropout in scRNA-seq data, a series of downstream analyses based on the gene expression will be affected, which may lead to the concealment of internal biological signals. Thus, the model-based imputation methods can correct the false zero expression and recover the original gene expression profile. MAGIC is a method based on manifold assumption, which deemed that the original high dimensional cell phenotype lies on a low dimensional manifold embedded within the measurement space. It utilized nearest neighbor graph representing manifold and then learned the underlying manifold via diffusion maps, restoring cellular phenotype back to the manifold and in the process realizing data imputation [9]. SAVER is a Bayesian-based approach to recover the true expression level of each gene in each cell, which modeled the gene count data and estimated the true expression as well as a posterior distribution quantifying the

uncertainty [8]. Both MAGIC and SAVER would potentially involve new biases into imputed data, since all gene expression levels were modified including those unaffected by dropouts. scImpute held that some of missing values may reflect true biological non-expression, which used a mixture model to learn the probability of each gen's dropout, and then imputed those dropout values with high probability in a cell by sharing information of other similar cells [7]. To our knowledge, DCA is the first imputation method combining deep learning with gene-specific distribution, which adopted the count distribution, namely negative binomial distribution with zero-inflation (ZINB), to capture gene-gene dependencies [16]. Despite these methods have achieved improvements in recovering potential biological information using imputation, they neglect the statistics characteristic of normalized scRNA-seq data for eliminating biological and technical biases.

## 3 Model selection

### 3.1 The analysis of scRNA-seq data

To elucidate the distribution characteristic of single cell dataset, we conduct data analysis on a publicly available real dataset from Klein et. al. For carrying out large-scale single-cell sequencing, they developed the inDrop platform which encapsulates cells into droplets with lysis buffer, reverse transcription (RT) reagents, and barcoded ologonu cleotide primers [26]. In addition, this platform can barcode the RNA from thousands of individual cells. They used this platform to study mouse embryonic stem cells, revealing in detail the population structure and the heterogeneous onset of differentiation after leukemia inhibitory factor (LIF) withdrawal. We use logarithmic normalization to preprocess the gene expression profile that consists of 2,717 cells and 24,175 genes, and filter genes that are unexpressed fewer than 10 cells form the original data, obtaining a total of 23,840 genes. As the typical feature of scRNA-seq data, the dropout events can be intuitively displayed by the histogram of the gene dropout ratio across cells in Fig. 1, whereby the number of genes with dropout ratio not less than 0.5 is 19,512, which indicates that a large number of genes are only expressed in a small number of cells.

Furthermore, we randomly select four genes to present the distribution characteristic of individual genes as shown in Fig. 2. It can be noted that the overall expression distribution of each gene has bimodal expression pattern, and the positive expression possesses a right-skewed characteristic. In order to further quantify the statistical characteristic, we approximate the skewness value of the population described below:

$$\gamma = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right], \tag{1}$$

where $\mu$ and $\sigma$ are the mean and standard deviation of the corresponding distribution, respectively. According to the moment estimator of the sample, the skewness values of the positive expression of the four genes are 1.49, 1.69, 1.61, and

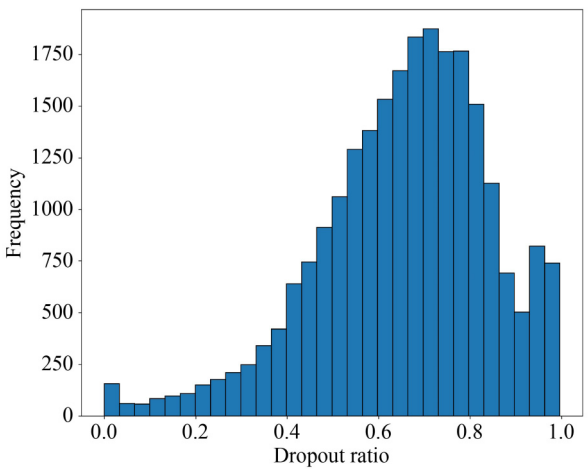



**Fig. 1** Histogram of dropout ratio of the 23,840 genes across the 2,717 cells in Klein dataset

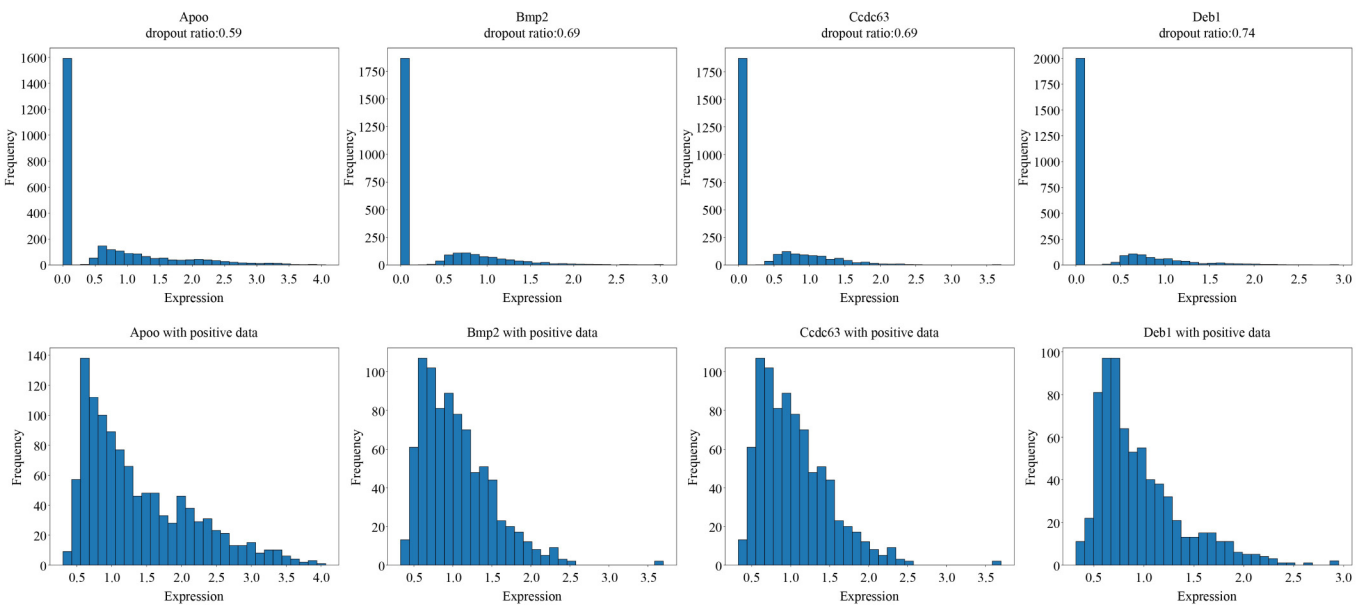



**Fig. 2** Histograms of the overall and positive expression distribution for Apoo, Bmp2, Ccdc63, and Deb1 genes in Klein dataset

1.96, respectively, which demonstrates the expression distribution of these genes are right-skewed. We further calculate the proportion of right deviation for the 23,840 genes, and the results show that the skewness values of the 99% genes are positive.

From the above analysis, we conclude that there are two typical characteristics of gene expression existed in the normalized scRNA-seq data. One is the bimodal expression mode as a whole, and the another is the right-skewness of positive expression. Both characteristics should be accounted in the modeling of scRNA-seq data in order to improve the accuracy of the downstream analysis.

### 3.2 The modeling of gene expression of normalized scRNA-seq data

Based on the bimodal expression mode of normalized scRNA-seq data, the two-part model is a reasonable choice, which generates zero data and positive data with two generating processes. The probability density of the two-part model is as follows:

$$f_{TP}(x \mid \pi, \boldsymbol{\theta}) = \pi I_{[x=0]} + (1-\pi) I_{[x>0]} f_p(x \mid \boldsymbol{\theta}), \tag{2}$$

where $\pi \in [0,1]$ is a parameter of Bernoulli distribution to fit for the probability of observing a positive-versus-zero expression and $I$ is the indicator function. If the condition in the subscript is satisfied, the value of the indicator function is 1.

For the skewed characteristic of positive expression, we adopt gamma and log-normal distributions as two alternatives to fit such random variables and the corresponding PDFs are as follows, respectively:

$$f_G(x \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x>0, \alpha>0, \beta>0, \tag{3}$$

$$f_{LN}(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi} x \sigma} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, \quad x>0, \sigma>0, \mu \in R. \tag{4}$$

To facilitate model comparison, we take the following assumptions for the two alternatives models, the two-part gamma model (TPGM) and the two-part log-normal model (TPLNM), respectively:

$$\begin{aligned} x_1 &\sim f_{TPGM}(\pi_1, \alpha, \beta), \\ x_2 &\sim f_{TPLNM}(\pi_2, \mu, \sigma). \end{aligned} \tag{5}$$

The distributions of the two models are as follows:

$$f_{TPGM}(x_1 \mid \pi_1, \alpha, \beta) = \pi_1 I_{[x_1=0]} + (1-\pi_1) I_{[x_1>0]} f_G(x_1 \mid \alpha, \beta), \tag{6}$$

$$f_{TPLNM}(x_2 \mid \pi_2, \mu, \sigma) = \pi_2 I_{[x_2=0]} + (1-\pi_2) f_{LN}(x_2 \mid \mu, \sigma) I_{[x_2>0]}. \tag{7}$$

Assuming the total number of cells is $n$, the log-likelihood functions of TPGM and TPLNM are respectively denoted as:

$$\begin{aligned} &\log(f_{TPGM}(x_1 \mid \pi_1, \alpha, \beta)) = \log\left(\prod_{i=1}^{n} f_{TPGM}\left(x_1^i \mid \pi_1, \alpha, \beta\right)\right) \\ &= n_0 \log \pi_1 + (n-n_0)\log(1-\pi_1) + (n-n_0)\alpha \log \beta \\ &-(n-n_0)\log\Gamma(\alpha) - (n-n_0)\beta\overline{x_1} + (n-n_0)(\alpha-1)\overline{\log x_1}, \end{aligned} \tag{8}$$

$$\begin{aligned} &\log(f_{TPLNM}(x_2 \mid \pi_2, \mu, \sigma)) = \log\left(\prod_{i=1}^{n} f_{TPLNM}\left(x_2^i \mid \pi_2, \mu, \sigma\right)\right) \\ &= n_0 \log \pi_2 + (n-n_0)\log(1-\pi_2) + (n-n_0)\log(\sqrt{2\pi}\sigma) \\ &-\sum_{i=1}^{n-n_0}\left[\log x_2^i + \frac{\left(\ln x_2^i - \mu\right)^2}{2\sigma^2}\right], \end{aligned} \tag{9}$$

where $n_0$ represents the number of dropouts of a specific gene, $\overline{x_1}$ denotes the mean of positive expression values, and $\overline{\log(x_1)}$ is the mean of logarithmic of positive expression values. Then the solution of parameters in Eqs. (8) and (9) can be achieved by applying the maximum likelihood estimation (MLE).

For TPGM, the closed-form solutions of $\pi_1$ and $\beta$ are respectively:

$$\widehat{\pi}_1 = \frac{n_0}{n}, \quad \widehat{\beta} = \frac{\alpha}{\bar{x}_1}. \tag{10}$$

The parameter $\alpha$ in Eq. (8) is updated iteratively using Newton's non-quadratic variation method [27]. The update equation of $\alpha$ is described as follows:

$$\frac{1}{\alpha_{t+1}} = \frac{1}{\alpha_t} + \frac{\log(\alpha_t) - \psi(\alpha_t) + \overline{\log x_1} - \log \overline{x_1}}{\alpha_t^2\left(\frac{1}{\alpha_t} - \psi'(\alpha_t)\right)}, \tag{11}$$

where $\psi'(\cdot)$ is the first derivative of the digamma function $\psi(\cdot)$. The above calculation is repeated until the convergence of $\alpha$ or a maximum number of iterations is reached.

For TPLNM, the closed-form solutions of $\pi_2$, $\mu$, and $\sigma$ are respectively:

$$\begin{aligned} \widehat{\pi}_2 &= \frac{n_0}{n}, \quad \widehat{\mu} = \frac{\sum_{i=1}^{n-n_0} \log x_2^i}{n-n_0}, \\ \widehat{\sigma}^2 &= \frac{\sum_{i=1}^{n-n_0}\left(\log x_2^i - \frac{\sum_{i=1}^{n-n_0} \log x_2^i}{n-n_0}\right)^2}{n-n_0}. \end{aligned} \tag{12}$$

In order to investigate the rationality of the two alternatives, the two-sample Kolmogorov-Smirnov (KS) test is applied to the total 23,840 genes, and the results show that about 95% and 92% genes are well fitted by TPGM and TPLNM, respectively (p-value $\geqslant 5\%$), which indicates that both of the models can fit the gene expression well. Specifically, the Cpe and Hsd11b1 genes are randomly selected for displaying the goodness-of-fit of the two models and Table 1 describes the empirical statistical information for the expression of both genes. According to the MLE solutions of the two models, the statistics including estimated parameters, mean, variance, skewness, log-likelihood, Akaike Information Criterion (AIC) and p-value of the KS tests, are summarized in Table 2. It is observed that the first-order and second-order statistics obtained by the two models are approximately consistent with the corresponding empirical statistics of the two genes. Though the KS test results show that both genes are well

**Table 1** Relevant expression statistics of the expression of Cpe and Hsd11b1 in Klein dataset

| Gene | Min | Max | Mean | Variance | Zero ratio | Skewness |
|---|---|---|---|---|---|---|
| Cpe | 0 | 3.51 | 0.62 | 0.56 | 0.53 | 0.57 |
| Hsd11b1 | 0 | 2.56 | 0.17 | 0.15 | 0.81 | 1.53 |

**Table 2** Results obtained from the MLE of TPGM and TPLNM for the expression level of Cpe and Hsd11b1 in Klein dataset

| Two-part models | Gene | MLE estimation | $\widehat{mean}$ | $\widehat{var}$ | $\widehat{Skewness}$ | KS-test | Log-likelihood | AIC |
|---|---|---|---|---|---|---|---|---|
| TPGM | Cpe | $\widehat{\pi}_1 = 0.53$<br>$\widehat{\alpha} = 6.40$<br>$\widehat{\beta} = 4.86$ | 0.62 | 0.56 | 0.79 | 0.97 | −2787.02 | 5580.04 |
| | Hsd11b1 | $\widehat{\pi}_1 = 0.81$<br>$\widehat{\alpha} = 7.31$<br>$\widehat{\beta} = 8.29$ | 0.17 | 0.14 | 0.74 | 0.80 | −1460.48 | 2936.96 |
| TPLNM | Cpe | $\widehat{\pi}_2 = 0.53$<br>$\widehat{\mu} = 0.20$<br>$\widehat{\sigma} = 0.41$ | 0.62 | 0.59 | 1.36 | 0.25 | −2803.16 | 5612.31 |
| | Hsd11b1 | $\widehat{\pi}_2 = 0.81$<br>$\widehat{\mu} = -0.20$<br>$\widehat{\sigma} = 0.37$ | 0.17 | 0.14 | 1.19 | 1.00 | −1442.60 | 2891.20 |

modeled by the two models, the values of AIC indicate that Cpe tends to select TPGM and Hsd11b1 favors TPLNM. To intuitively display the goodness-of-fit of the two models, the fitting curves of the two distributions corresponding to the two genes are shown in Fig. 3.

In addition, we randomly select five batches of genes with two hundred genes per batch from the 23,840 genes, and calculate the percentage of different model selections according to the AIC values. The results shown in Fig. 4 indicates that TPLNM fits more genes than TPGM.

The analysis confirms that the two-part model with the two right-skewed distributions can capture the bimodal expression mode of normalized scRNA-seq data. Therefore, gamma and lognormal can be used as alternative models to fit positive right-skewed expression.

## 4 Autoencoder based on a two-part-generalized-gamma distribution

In this section, we first introduce Two-Part-Generalized-Gamma (TPGG) distribution preparing for building AE-TPGG and then give a detailed description of the personalized autoencoder, AE-TPGG, for scRNA-seq data analysis.

### 4.1 TPGG distribution

In Section 2, we have testified that TPGM and TPLNM can be used as two alternatives for modeling gene expression of scRNA-seq data. However, model selection is a necessary step in order to obtain an appropriate model for each gene, which brings the extra computational cost, especially for large-scale datasets. To achieve adaptive model selection, we select TPGG model for fitting gene expression, making use of the advantages of its powerful fitting ability, which flexibly selects appropriate distribution by adjusting parameters. Therefore, we assume gene expression follows TPGG distribution that consists of two components: a point mass at zero that captures the high dropout events and a generalized-gamma distribution modeling positive right-skewed expression. The TPGG distribution is parameterized with shape and scale parameters of the generalized gamma distribution ($\alpha$, $\beta$, and $\gamma$) and $\pi$ that is the parameter of Bernoulli distribution fitting for the probability of a positive-versus-zero outcome, denoted as:

$$f_{TPGG}(x|\pi,\alpha,\beta,\gamma) = \pi I_{[x=0]} + (1-\pi) I_{[x>0]} f_{GG}(x|\alpha,\beta,\gamma), \tag{13}$$

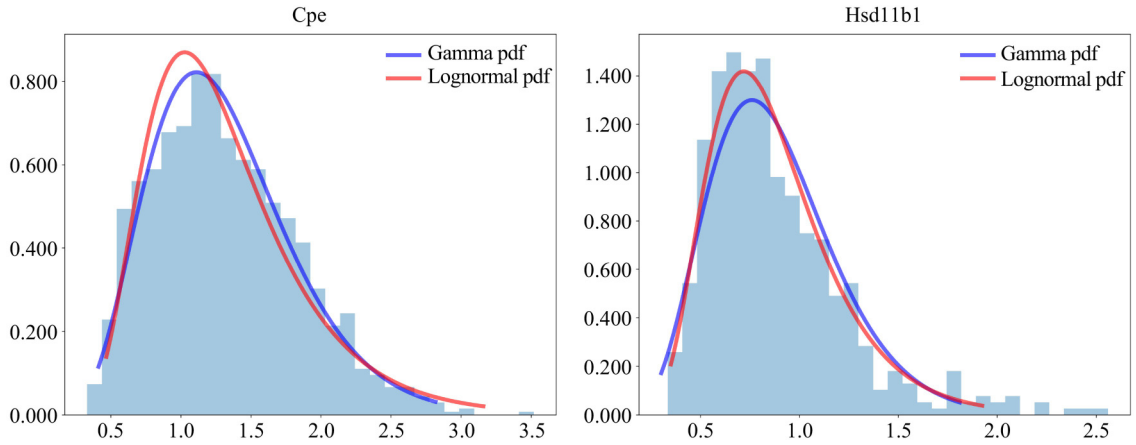



**Fig. 3** The sample distributions and estimated gamma distributions of the positive expression values of Cpe and Hsd11b1 in Klein dataset

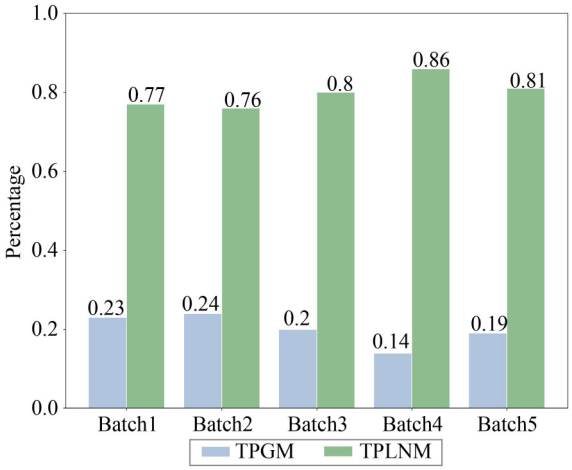



**Fig. 4** The percentage of different model selections according to the AIC values for five batched of randomly selected genes from Klein dataset

$$f_{GG}(x|\alpha,\beta,\gamma)=\frac{\gamma}{\Gamma(\beta)}\frac{x^{\beta\gamma-1}}{\alpha^{\beta\gamma}}e^{-(\frac{x}{\alpha})^{\gamma}}, \tag{14}$$

where $\pi\in[0,1]$, $\alpha>0$, $\beta>0$, and $\gamma>0$. $\Gamma(\cdot)$ is a gamma function. As an appealing feature, TPGG is flexible in that it encompasses several well-known subfamilies. For instance, TPGG becomes TPGM in case of $\gamma=1$ and TPLNM is also obtained as a limiting case when $\beta\to\infty$. Therefore, the proposed model can adaptively select two alternative models by adjusting its own parameters to fit the scRNA-seq data better.

### 4.2 AE-TPGG architecture

In addition to the distribution characteristic modeled by TPGG, AE-TPGG also accounts for the intrinsic relationship between genes. Under the assumption that gene expression follows the TPGG distribution, we use the autoencoder to estimate the four parameters of the distribution conditioned on each cell data for each gene. Thus, unlike traditional autoencoder where one input corresponds to one output, dropout($\pi$), shape($\alpha$,$\beta$) and scale($\gamma$) parameters are set as the decoder outputs. Each atomic unit consisting of the four outputs corresponds to the parameters of the TPGG distribution for a single gene, given as $\alpha$, $\beta$, $\gamma$, and $\pi$. According to this design, the input layer and four output layers corresponding to these parameters of single cell are of the same size. Furthermore, unlike regular autoencoders using MSE as loss, we adopt the form of the likelihood function associating the four outputs representing the four parameters with the original input of each gene, to optimize the model parameters. Meanwhile, low dimensional representation can be obtained by the encoder in the process of maximizing likelihood function and this embedded representation reflects the inherent characteristic of expression data for each cell. The architecture of AE-TPGG is depicted in Fig. 5 and the rules of forward propagation are given below:

$$\begin{aligned}
&H_l=\sigma(H_{l-1}W_{l-1}),\ (l=1,\ldots,D-1),\\
&\Pi=sigmoid(H_{D-1}W_{D_\pi}),\\
&A=softplus(H_{D-1}W_{D_\alpha}),\\
&B=softplus(H_{D-1}W_{D_\beta}),\\
&R=softplus(H_{D-1}W_{D_\gamma}).
\end{aligned} \tag{15}$$

The first line expression in Eq. (15) describes the forward propagation process of the model before decoding the outputs, where $D-1$ is the index of the penultimate layer of the network. $H_0$ represents an $mxn$ input matrix $X$ that is the normalized expression profile, where $m$ and $n$ correspond to the number of cells and genes, respectively. The outputs of the model are four inferred parameter matrices, i.e., $\Pi$, $A$, $B$, and $R$, and each size of them is consistent with that of the input expression profile $X$. $\sigma(\cdot)$ denotes the activation function acting on each element of the data matrix, such as $ReLU(x)=\max(0,x)$ or $sigmoid(x)=\frac{1}{1+e^{-x}}$. The remaining expressions are multiple decoded outputs, where $\Pi,A,B$, and

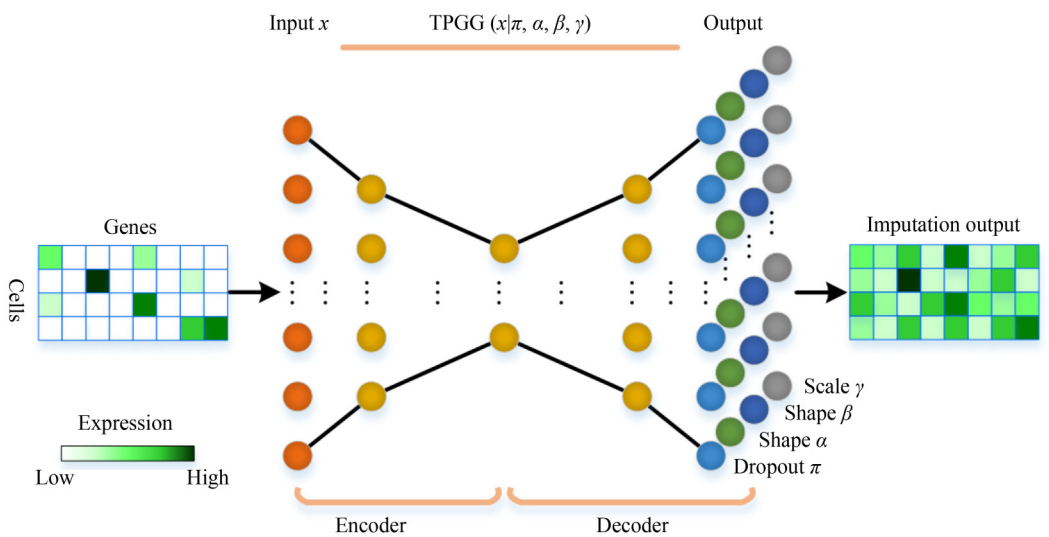



**Fig. 5** The framework of the proposed AE-TPGG. First, the encoder module of AE-TPGG automatically extracts a high-level compressed representation of the gene expression profile based on the multiple full connection layers. Subsequently, the decoder component of AE-TPGG derives the parameters $\alpha$, $\beta$, $\gamma$ and $\pi$ to acquire the imputation output achieved by the expectation calculation of each gene expression level. The input is associated with the output by optimizing the negative log-likelihood of the TPGG shown at the top of the diagram

$R$ are the matrix representations corresponding to the parameters of TPGG model for all genes. We apply distinct types of activation functions to the decoding output layers according to the feasible range of the parameters of TPGG distribution. For shape and scale parameters, the values are positive and thus the corresponding decoder outputs are transformed into positive value using softplus function, which is $softplus(x) = \log(1 + e^x)$. The values of $\Pi$ are limited to the range between zero and one, hence the sigmoid function is a suitable choice to impose the conversion on this decoder layer. The symbols $W$ with different subscripts correspond to the connection weight matrix between front and back layers.

### 4.3 AE-TPGG loss

Here, we treat the negative log-likelihood of the TPGG as loss, effectively connecting the input and output. Meanwhile, an adjustable regularization term is added to prevent the influence of static noise on the optimization objective and the irrelevant components of learnable parameters. The loss of AE-TPGG is represented as:

$$\begin{aligned} J(W_{\{S\}}) &= -\log(L_{TPGG}(X|\Pi, A, B, R)) + \lambda \sum_{t\in\{S\}} \|W_t\|_F^2 \\ &= -\log\left(\prod_{i=1}^{m}\prod_{j=1}^{n} TPGG(x_{ij}|\pi_{ij}, \alpha_{ij}, \beta_{ij}, \gamma_{ij})\right) + \lambda \sum_{t\in\{S\}} \|W_t\|_F^2 \\ &= -\sum_{i=1}^{m}\sum_{j=1}^{n} \log(TPGG(x_{ij}|\pi_{ij}, \alpha_{ij}, \beta_{ij}, \gamma_{ij})) + \lambda \sum_{t\in\{S\}} \|W_t\|_F^2, \end{aligned} \tag{16}$$

where $\{S\}$ denotes $\{0, \ldots, D-2, D_\pi, D_\alpha, D_\beta, D_\gamma\}$. Here, $x_{ij}$ represents the expression level of gene $i$ on cell $j$. The $\lambda$ is a hyperparameter which balances the loss of negative logarithmic likelihood of TPGG and weight penalty terms, and $\|\cdot\|_F$ denotes the Frobenius norm.

### 4.4 Imputation of scRNA-seq data

The imputation value is generated with the expectation of the estimated TPGG distribution, which is defined below:

$$\begin{aligned} E[x_{ij}] &= (1 - \widehat{\pi}_{ij}) E[x_{ij}|x_{ij} > 0] \\ &= (1 - \widehat{\pi}_{ij}) \frac{\widehat{\alpha}_{ij}\Gamma\left(\widehat{\beta}_{ij} + \frac{1}{\widehat{\gamma}_{ij}}\right)}{\Gamma(\widehat{\beta}_{ij})}, \end{aligned} \tag{17}$$

where $\widehat{\pi}_{ij}$, $\widehat{\alpha}_{ij}$, $\widehat{\beta}_{ij}$, and $\widehat{\gamma}_{ij}$ are the estimated parameters of each gene from the AE-TPGG network.

### 4.5 Implementation

We utilize Keras [28] and its TensorFlow [29] backend to implement our model. RMSProp method is adopted for the optimization with initial learning rate, and the learning rate is reduced by multiplying 0.1 if the validation loss does not change with a fixed number of epochs. We adopt the dropout technology to prevent the model from over-fitting. For the size of the hidden layers, the bottleneck layer has 32 neurons and the number of neurons for other hidden layers are set to 64 for default settings of three hidden layers. Meanwhile, for the sake of flexibility, several different network structures are optional according to the size and complexity of datasets. The implementation of AE-TPGG details are summarized in Algorithm 1. The source code is available at Github.com/PUGEA/AE-TPGG website.

**Algorithm 1** The training procedure for AE-TPGG.

**Require:** The normalized expression profile of scRNA-seq data: $X$; the number of hidden layers and the neuron nodes of each layer: $(h_1, h_2, h_3)$; initial learning rate: $\alpha_l$; the parameter for regularization term: $\lambda$; mini-batch size: $m$; training iterations: $I$; and the ratio of validation set: $r$;

**Ensure:** Low dimensional representation and imputation outputs: $H$ and $\widehat{X}$ ;

Initialization $W_S$ using uniform distribution or truncated normal distribution sampling;
**for** each $t \in [1, I_t]$ **do**
  **for** each $tt \in [1, \lceil n_{sample\_size}/n_{minibatch\_size} \rceil]$ **do**
    Do forward propagation to obtain parameters of TPGG distribution according to Eq. (15);
    **for** each $x_{ij}$ **do**
      **if** $x_{ij} = 0$ **then**
        $log\_likelihood\_value = \log\widehat{\pi}_{ij}$;
      **else**
        $log\_likelihood\_value = \log(1 - \widehat{\pi}_{ij}) +$
        $\log\gamma_{ij} + (\widehat{\beta}_{ij}\widehat{\gamma}_{ij} - 1)\log x_{ij} - \log(\Gamma(\widehat{\beta}_{ij}))$
        $-\widehat{\beta}_{ij}\widehat{\gamma}_{ij}\log\widehat{\alpha}_{ij} - (\frac{x_{ij}}{\widehat{\alpha}_{ij}})^{\widehat{\gamma}_{ij}}$;
      **end if**
    **end for**
    Sum the log likelihood of minibatch cells
    $J_{minibatch}$;
    Calculate the loss of regularization term $J_r$;
    Update the network parameters with gradient by
    $(-J_{minibatch} + J_r)$;
  **end for**
**end for**
Do forward propagation to obtain $H$;
Calculate $\widehat{X}$ according to Eq. (17);
**return** $H, \widehat{X}$;

## 5 Experiments results

We demonstrate the performance of the proposed method on several biological problems involving real scRNA-seq datasets.

### 5.1 Improvement in capturing cell population structure in real data using imputation data

It is well known that cell heterogeneity is the basic feature of organisms, so the recognition of cell subtypes is an important task of single cell data analysis. In this section, we study the influence of the imputation data obtained from our proposed method on the recognition of phenotypic information of real single cell data and compare with other mainstream methods.

#### 5.1.1 Dataset description and experimental setup

We perform the validation of AE-TPGG on four publicly available scRNA-seq datasets with different cell numbers. The Klein dataset described in Section 3.1 is also used here, and the other three are described below:

- **Deng dataset** This dataset is derived from the single cell RNA sequencing of preimplantation cells of mouse embryos in mixed background to study the expression of alleles [30]. It is found that a large number of autosomal genes have monoallelic expression, and the expression of two alleles occurs independently. The

expression of single allele is random because of the great difference in the embryonic cells of close relatives. The results indicate that abundant random monoallelic expression is emerged from independent and stochastic allelic transcription in mammalian cells. They adopt vivo F1 embryos to isolate single cells, and Smart-seq protocol to obtain transcriptome expression profiles. Cell phenotypes are labeled by different embryonic and developmental status.

- **Kolodziejczyk dataset** The culture conditions of embryonic stem cells play an important role in maintaining long-term self-renewal and affect their pluripotent state [31]. Kolodziejczyk et. al. used full-transcript single-cell RNA sequencing of mESCs cultured in three different conditions: serum, 2i, and the alternative ground state a2i. They find that overall levels of intercellular heterogeneity are comparable across the three conditions, but different sets of genes are variably expressed.
- **pbmc1-10Xv2 dataset** To systematically and comprehensively evaluate the scale and capabilities of single-cell RNA-sequencing methods, Ding et. al. developed a flexible computational pipeline named scumi to compare seven methods for single-cell and/or single-nucleus profiling on three types of samples, cell lines, peripheral blood mononuclear cells and brain tissue [32]. The pbmc1-10Xv2 dataset is from human peripheral blood mononuclear cells (PBMCs) performed on the high-throughput method, the 10x Chromium (v2) platform, which consists of nine cell types that are B cell, CD14+ monocyte, CD16+ monocyte, CD4+ T cell, Cytotoxic T cell, Dendritic cell, Megakaryocyte, Natural killer cell and Plasmacytoid dendritic cell.

The statistical information of the four datasets is summarized in Table 3, and the high dropout ratio is a common feature of the four datasets.

For making a reasonable evaluation, we use the same experimental setup for various imputation methods. In order to alleviate the curse of dimensionality and reduce the computation cost of learning tasks, the feature selection is applied to all datasets for picking up the highly variable genes (HVGs). Specifically, the gene counts for each cell are divided by the total counts for that cell and multiplied by the scale factor. Then, the natural-log transformation is used to prevent a few large observations from being extremely influential. Finally, the first 5,000 HVGs are retained according to the variability of the genes. After that, we perform visualization and clustering to evaluate various imputation methods on recognition of cell types for the four datasets.

The baseline imputation methods we compare are four mainstream approaches, including MAGIC, DCA, SAVER, and scImpute, and an overview of these methods is described in Section 2.

#### 5.1.2 Evaluation measures

In order to evaluate the clustering results comprehensively, we employ four metrics, Accuracy (ACC), Adjusted Rand Index (ARI), Normalized Mutual Information (NMI) and F1-score (F1), which are defined respectively as follows:

$$\begin{aligned} ACC &= \frac{1}{N}\sum_{i=1}^{N} 1(\widehat{y_i} = y_i), \\ ARI &= \frac{RI - E[RI]}{\max(RI) - E[RI]}, \\ NMI &= 2 \cdot \frac{MI(U,V)}{H(U) + H(V)}, \\ F1 &= 2 \cdot \frac{precision \cdot recall}{precision + recall}, \end{aligned} \tag{18}$$

where $\widehat{y_i}$ is the predicted label, $y_i$ is the ground-truth label, $1(\cdot)$ is an indicator function, $RI$ is the Rand index, $U$ and $V$ represent divisions of true and predictive labels, $MI(\cdot)$ denotes the mutual information, and $H(\cdot)$ is entropy. The larger values of these metrics means higher concordance between the predictions and the truth.

#### 5.1.3 Experimental results of the four datasets

In this section, we will compare the performance of different imputation methods through discovering the structural information of scRNA-seq data via visualization and clustering. In addition, the scRNA-seq data forms include two categories, count data and normalized data. Therefore, we further investigate the influence of data form towards the recognition of cell types. Besides, since the cell-cell distance matrix in MAGIC is based on Euclidian distance, the added imputation method named MAGIC-C is based on count data to observe the influence of data form on imputation. To facilitate comparison, the original MAGIC method based on the normalized data is denoted as MAGIC-N to easily distinguish. Thus, the objects we aim to study are expressed in the following eight forms, the raw count data with dropout denoted as With dropout-C, the raw normalized data with dropout denoted as With dropout-N, and the six imputed expression profiles using MAGIC-C, MAGIC-N, DCA, SAVER, scImpute and AE-TPGG, which are applied to the results display of Figs. 6 and 7.

In visualization experiment, we first project the original data to the first 50 principal component directions by principal component analysis (PCA), and then use t-distributed logistic neighbor embedding (t-SNE) to visualize on a two-

**Table 3** Statistics of the four real datasets

| Datesets | Sequencing protocol | Cell types | # of cells | # of genes | Zero ratio |
|---|---|---|---|---|---|
| Deng | Smart-seq | 10 | 268 | 22,431 | 0.60 |
| Kolodziejczyk | SMARTer | 3 | 704 | 38,616 | 0.71 |
| Klein | inDrop | 4 | 2,717 | 24,175 | 0.66 |
| pbmc1-10Xv2 | 10x Chromium (v2) | 9 | 6,444 | 22,280 | 0.96 |

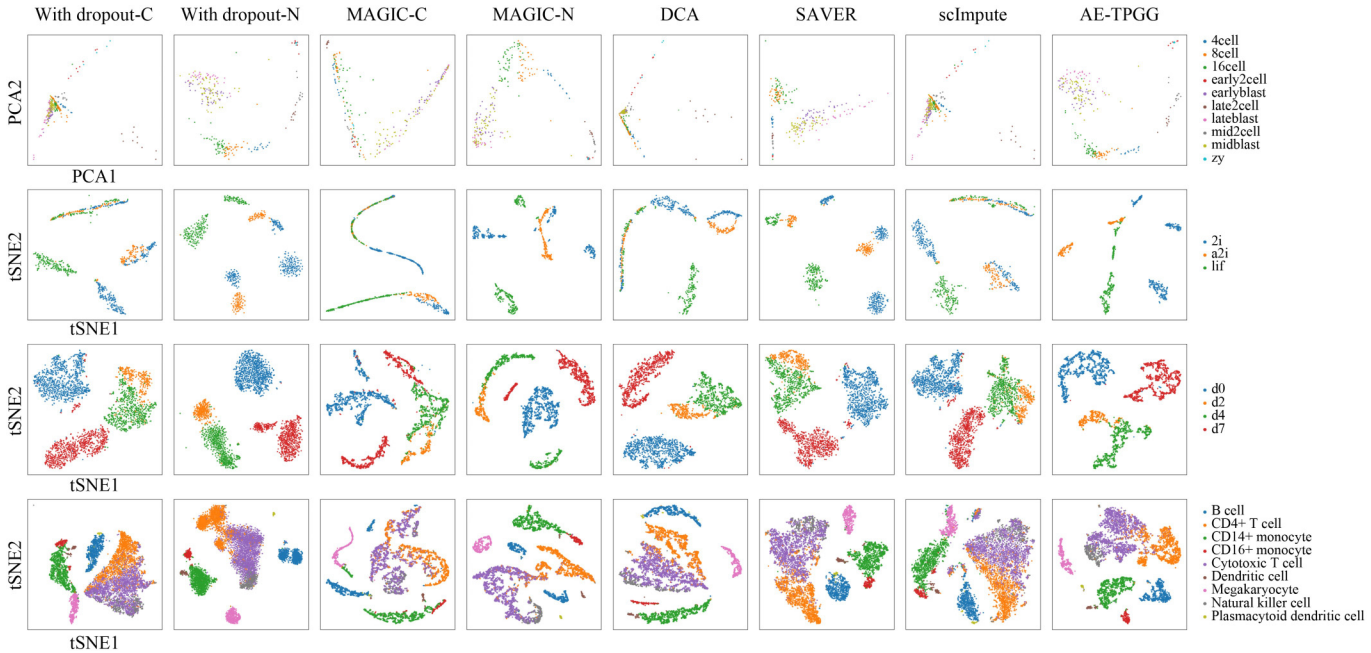



**Fig. 6** The two-dimensional visualization of Deng dataset (1st row), Kolodziejczyk dataset (2nd row), Klein dataset (3rd row), pbmc1-10Xv2 dataset (4th row). Columns correspond to the raw data of With dropout-C and With dropout-N, the imputed data using MAGIC-C, MAGIC-N, DCA, SAVER, scImpute and AE-TPGG. Cells are colored according to cell types

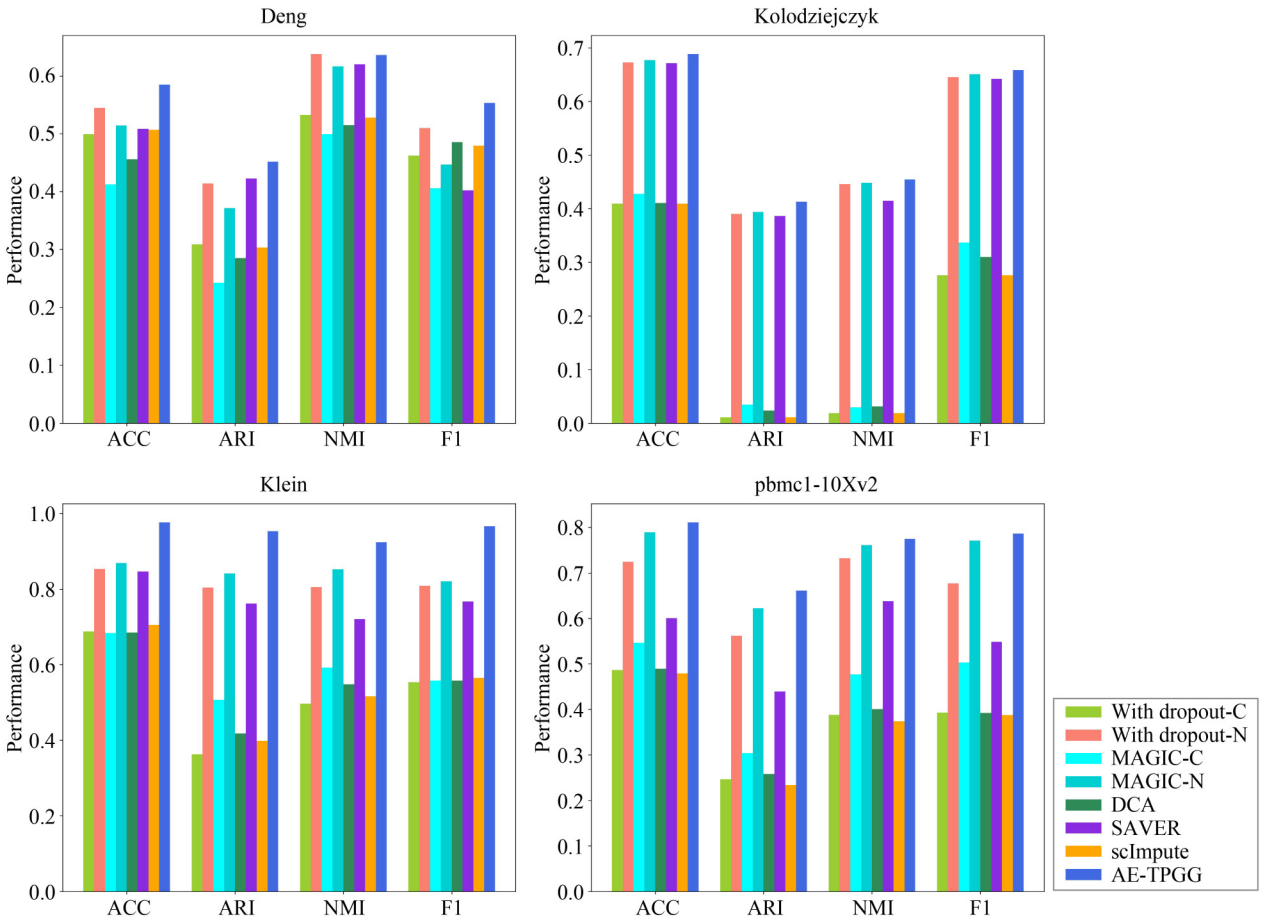



**Fig. 7** The clustering performance of the raw data of With dropout-C and With dropout-N, and the imputed data using MAGIC-C, MAGIC-N, DCA, SAVER, scImpute and AE-TPGG on the Deng dataset, Kolodziejczyk dataset, Klein dataset and pbmc1-10Xv2 dataset measured by ACC, ARI, NMI and F1

dimensional plane. Owning to the similarity between cells for the Deng dataset, the low dimensional representation obtained by t-NSE is too crowded, so we directly project it to two-dimensional plane by PCA. The cells are labeled according to the cell types provided by the original datasets. The visualization results of the four datasets are shown in Fig. 6. It can be seen that With dropout-N can significantly improve the discovery of structural information compared with With dropout-C. From comparison of MAGIC-C and MAGIC-N, the imputation of normalized data produces clearer the structural information than that of count data. Notice that the imputation based on normalized data using AE-TPGG has certain advantages over the imputation data based on count data using DCA in structure information discovery, although the two methods adopt the same method of parameter estimation. Thus, the visualization results indicate the necessity of data normalization in the preprocessing of scRNA-seq data, which can eliminate the biological and technical deviations to a certain extent and contribute to the discovery of structure information of scRNA-seq data. Although there are differences in imputation strategies among various methods, the imputation data produces more compact and smooth local structure on the whole.

We further perform clustering on the imputation data obtained from different methods, and measure the clustering performance using the four metrics in Eq. (18). To speed up the calculation, we reduce the original high-dimensional data to 50 dimensions through PCA, and then obtain the clusters by K-means. Considering clustering results that is sensitive to the random initialization of cluster centers, the k-means++ is utilized to initialize the centroids [33], and run the initialization for 10 times. The optimal result is finally selected in the sense of inertia. In order to obtain statistically meaningful evaluation results, each metric is measured by the mean value of the 10 runs. The clustering results of the four datasets are shown in Fig. 7. Compared with the raw count data denoted as With dropout-C, the adoption of With dropout-N significantly improves the clustering performance and enhance the recognition of phenotype information. This is especially obvious in comparison between MAGIC-C and MAGIC-N. All these imputation approaches account for data normalization at different stages in the models. MAGIC-N, scImpete and our method perform normalization at the preprocessing stage, while DCA and SAVER are statistical models based on count data and consider the normalization of gene expression in the process of parameter inference. In particular, DCA normalizes the input data in the middle hidden layer, and SAVER introduces the normalization factor at the final prediction of expression level, which are lack of distribution hypothesis of the normalization data. Although scImpute is based on the continuous normalized data, and ignores the typical bimodal expression pattern of scRNA-seq data. The experimental results show that it does not perform as well as other methods designed for normalized data. We speculate that the reason may due to the fact that the hyperparameter t that is used to determine whether the genes need to be imputed and the pre-set subgroup division are set manually, which are highly dependent on prior knowledge.

From the above comparison results, it can be noted that our method is superior to other imputation methods in clustering performance on the four real datasets. Particularly, the comparison between the methods based on count data with those based on normalized data indicates that the necessity of data normalization for phenotypic information recognition. Moreover, the adopted TPGG distribution, the deep autoencoder embedded in the model structure, and the parameter estimation approach make it possible to automatically capture the dependence between genes while inferring parameters, so as to achieve effective data imputation and improve the discovery of cell phenotype information in scRNA-seq data.

### 5.2 Improvement in capturing droplet-based purified cell types via low-dimensional representation

Dimensionality reduction plays a crucial role in analyzing high-dimensional scRNA-seq data, which is also the most typical application scenario of autoencoder. In this section, we conduct experiments on a real dataset to validate the performance of our model in dimension-reduction. Here, we choose expression profile of 68K PBMCs dataset based on droplet microfluidic control system, which consists of 68,579 cells and 20,387 genes, and cells have been divided into 10 purified cell subtypes [34].

We utilize SCANPY to select 1,000 highly variable genes, and the visualization results from the t-SNE method and the two-dimensional output of the bottleneck layer of AE-TPGG are displayed in Fig. 8. From Fig. 8, there is overwhelming evidence that heterogeneity exists in PBMCs, even though these cells stem from the same type. Meanwhile, due to the functional similarity of some cells, phenotypes of these cells are overlapped with low-dimensional representation. It is apparent that our model achieves the higher intra-class compactness than t-SNE. The results of the four clustering metrics are depicted in Fig. 9. It can be observed that AE-TPGG also obtains obviously higher clustering performance than t-SNE indicating AE-TPGG is able to discover the inherent low dimensional structure in the scRNA-seq data resulting in more accurate clustering results.

### 5.3 Improvements in protein and RNA co-expression analysis using imputation data

For large-scale simultaneous measurement of epitopes and transcriptomes in single cells, Stoeckius et. al. developed Cellular Indexing of Transcriptomes and Epitopes by sequencing (CITE-seq), which combined highly multiplexed antibody-based detection of protein markers together with unbiased transcriptome profiling for thousands of single cells in parallel [35]. The similar verification approach is adopted as DCA [16] and adopt levels of cell surface marker protein as ‘benchmark’, which have high abundance and less influence by dropout events. The original 8,617 cord blood mononuclear cells (CBMCs) are filtered with less than 90% human Unique Molecular Identifier (UMI) counts and 8,005 human cells are retained. In order to speed up the calculation, we selected the expression profile of the top 5,000 HVGs. To reveal the co-expression of gene and protein, we investigate the relationship between gene and protein levels according to their enrichment

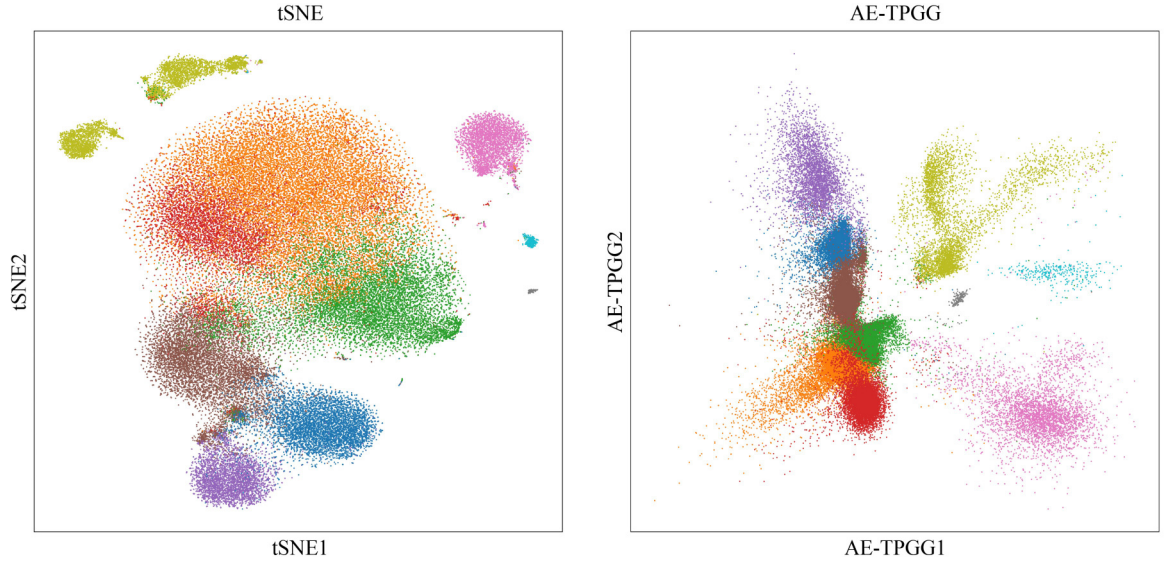



**Fig. 8** The t-SNE (left) and AE-TPGG (right) projections of 68K PBMCs, colored according to the 10 purified cell subtypes

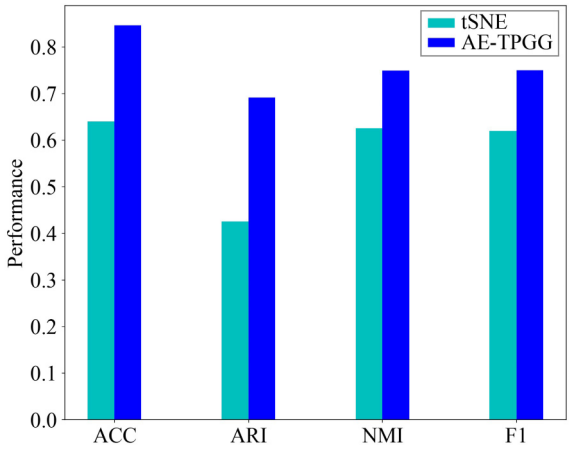



**Fig. 9** The Clustering performance of the PBMC 68K dataset measured by ACC, ARI, NMI and F1

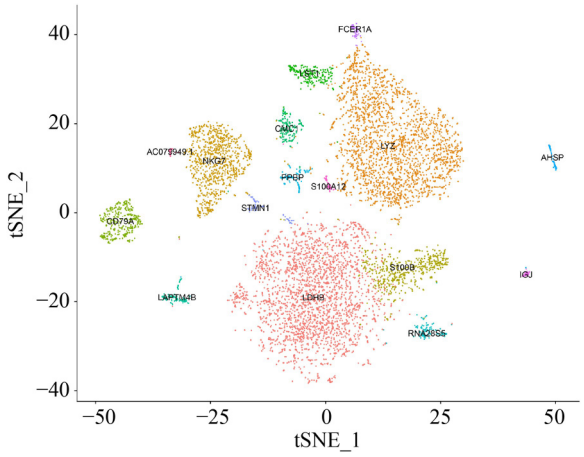



**Fig. 10** The t-SNE visualization of the transcriptomic profiles of cord blood mononuclear cells from Stoeckius. Cell types are labeled with maker genes

in cells. Cell subpopulations with graph-based clustering method are subsequently identified [36,37]. For constructing graph, the top 25 principal components by PCA are used to calculate k-nearest neighbors and construct a shared nearest neighbor (SNN) graph that captures the similarity between two nodes in terms of their connectivity in the neighborhood. In the next step, the dense subgraphs are found by recursively merging quasi-cliques in SNN graphs where the hyperparameter $r$, which is used to adjust the density of each cluster, is set to 0.5. Finally, the 16 cell types are obtained and the two-dimensional visualization is displayed in Fig. 10. The cell types are tagged according to the differential genes among these subgroups.

We visualize the six known marker proteins (CD3, CD8, CD56, CD16, CD11c, and CD14) and corresponding mRNAs (CD3E, CD8A, NCAM1, FCGR3A, ITGAX, and CD14) on RNA clusters where the first three genes and the rest are displayed in Fig. 11 and Fig. A1 of Appendix A, respectively. It is obvious that the original RNA and the co-expressed proteins present consistent enrichment areas in different subpopulations. However, due to the high dropout ratio of RNA data, RNA enrichment level in most subpopulations is obviously lower, especially for CD8A, NCAM1 and ITGAX which are only enriched in a small number of cells. This indicates the potential challenges exist in the analysis of cell heterogeneity only using the original transcriptome data. Therefore, data imputation is a feasible approach to alleviate the influence of dropout events. For better display, the imputed results of the first three genes from AE-TPGG and the baseline methods are shown in Figs. 11 and 12, and the imputed results of the other genes of all imputation methods are shown in Fig. A1 of Appendix A. The third panel of Fig. 11 and Fig. A1 in Appendix A shows the visualization results of imputed data obtained from AE-TPGG. It can be seen that both the enrichment regions and the enrichment level are highly consistent with those of the protein data. The same way is applied to visualize the imputed data of transcriptomic profile obtained from other imputation methods shown in Fig. 12 and Fig. A1 of Appendix A. It can be apparently found that the imputation data can significantly improve the co-

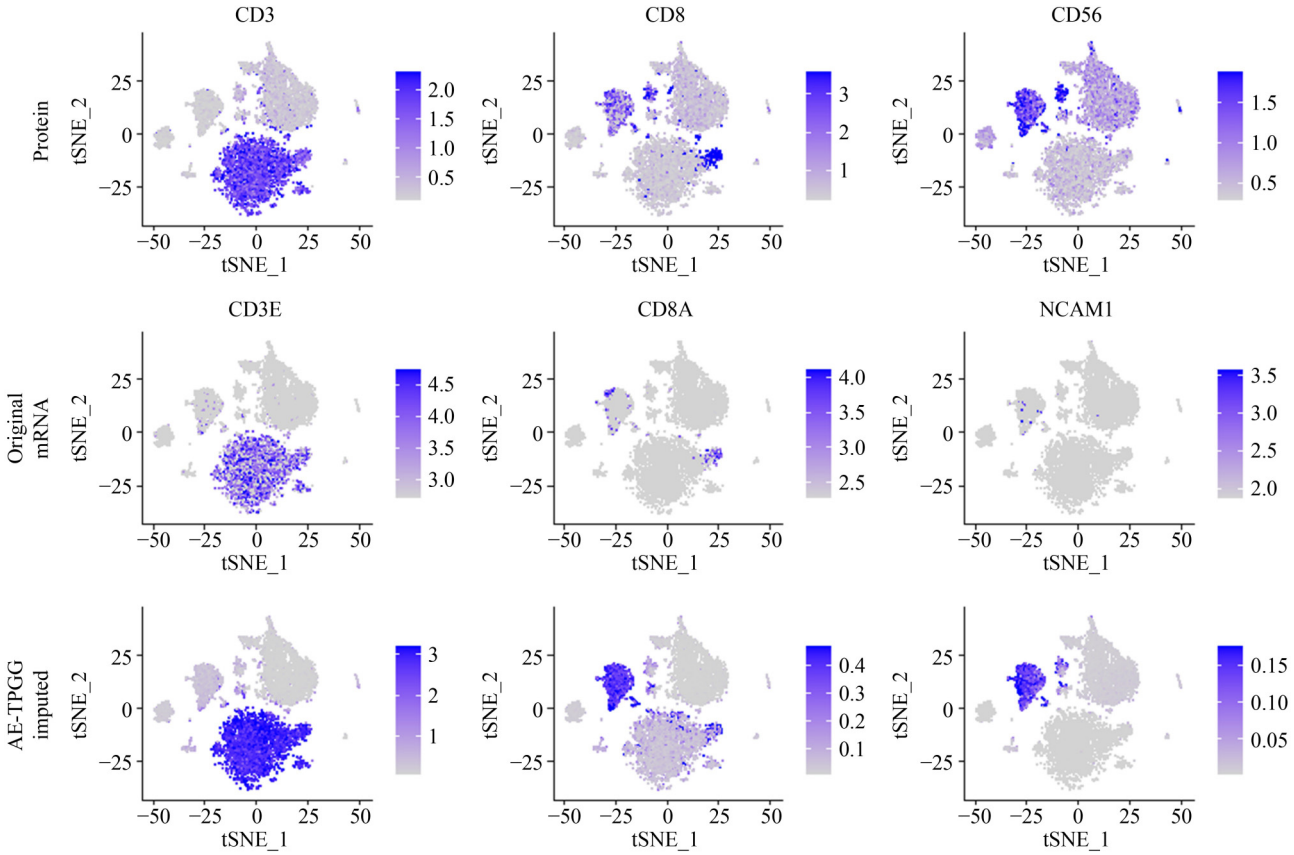



**Fig. 11** The t-SNE visualizations of the protein expression (1st row), RNA expression derived from the original data (2nd row), imputation data using AE-TPGG (3rd row). Columns correspond to CD3 (1st column), CD8 (2nd column), CD56 (3rd column) proteins and corresponding RNAs CD3E, CD8A and NCAM1

expression analysis between genes and proteins.

Especially, our method and MAGIC present better consistency with levels of the antibody-derived tags (ADT). In addition, the analysis of Spearman correlation is used to quantify the correlation between imputed data obtained from different methods and the corresponding surface proteins. The results depicted in Fig. 13 show that the correlation coefficients between the original mRNAs and the corresponding proteins are relatively low, while those between the imputed data and the proteins are significantly high. Therefore, AE-TPGG achieves competitive results compared with other imputation alternatives showing the superiority on improving protein and RNA co-expression analysis.

### 5.4 The recovery of time-course patterns using imputation data

In this section, we perform the experimental analysis on a semi-real single-cell dataset. Owing to dropout events in real single-cell datasets, the truth of the data is usually unknown. Therefore, we adopt a similar synthetic method for semi-real dataset as in MAGIC method. The validation set is established on bulk transcriptomic data measured using microarrays from 206 developmentally synchronized C. elegans young adults [38]. These samples are taken at regular time intervals during a 12-hour developmental time-course and consecutive samples, which present relatively small changes in expression. The dataset contains 206 samples and 15,855 informative genes. The generation method of single-cell dataset first exponentiates the original gene expression profile. Then the specific noise is added for each gene by subtracting random values sampled from an exponential distribution where the mean is calculated as the gene expression median multiplied by five and set all negative values to zero. Finally, the data is logarithmized and the final synthesized data contains 80% missing values. Figure 14 displays heatmaps of the top 200 highly variable genes that consists of 100 positive and 100 negative genes associated with time course within the dataset using expression data without dropout, with dropout, and the imputed data from various imputation methods. It can be seen that the time-varying gene expression pattern of the synthesized data is indistinct after introducing dropout values. After imputation by the AE-TPGG model, the gene expression pattern clearly restores compared with the validation set. Meanwhile, the imputed outputs of the four baseline methods described in Section 5.1.1 are visualized with the same experimental settings. As shown in Fig. 14, the imputation is able to recover part of expression pattern compared with the validation data, but our method has higher consistency than other methods and significantly alleviates the imputation deviations, including over-imputation and under-imputation.

To further compare with different imputation methods, we select the top 500 genes in validation set that are significantly related to the development and measure Person's correlation coefficients between gene expression and the known developmental pattern to qualitatively assess the pattern

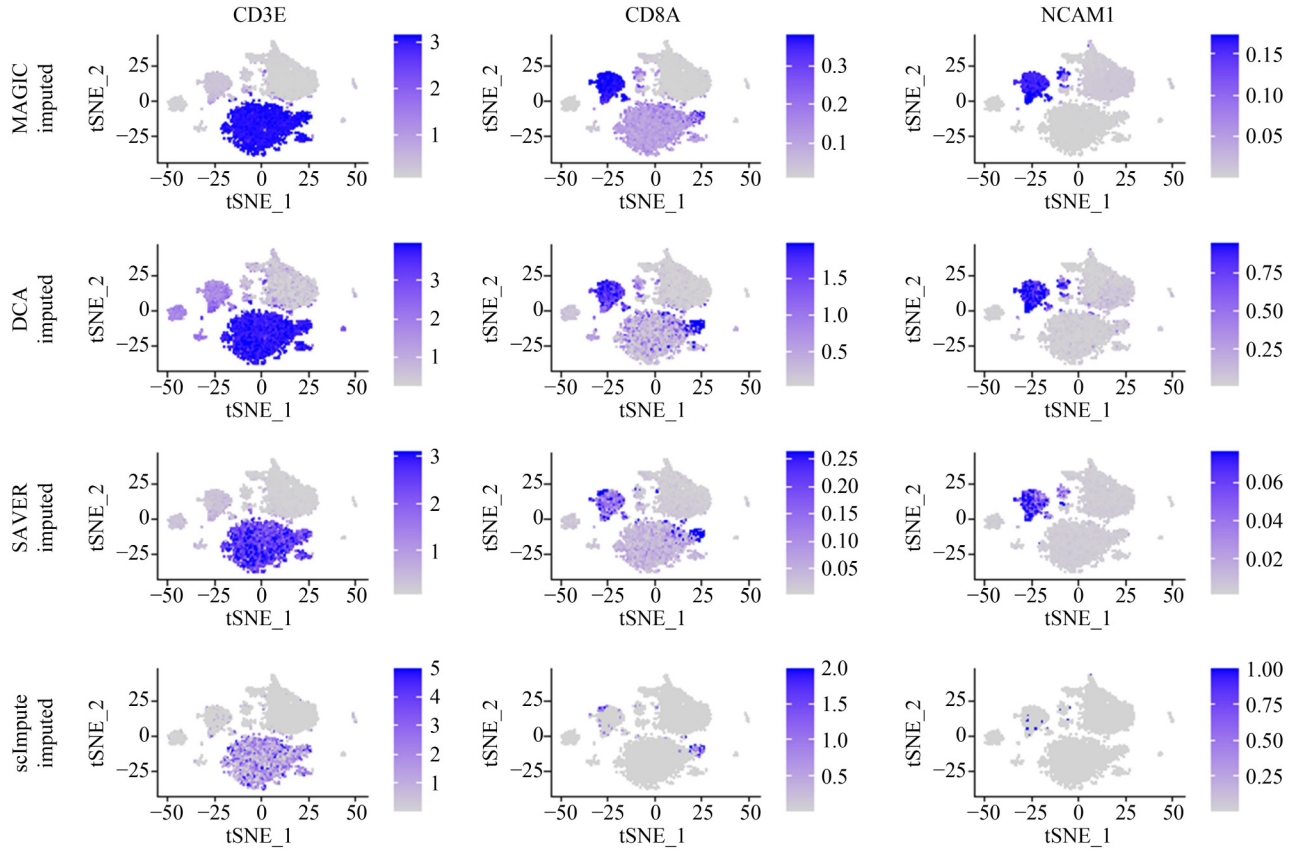



**Fig. 12** The t-SNE visualizations of imputation data of RNA expression using MAGIC (1st row), DCA (2nd row), SAVER (3rd row), and scImpute (4th row). Columns correspond to RNAs CD3E (1st column), CD8A (2nd column) and NCAM1 (3rd column)

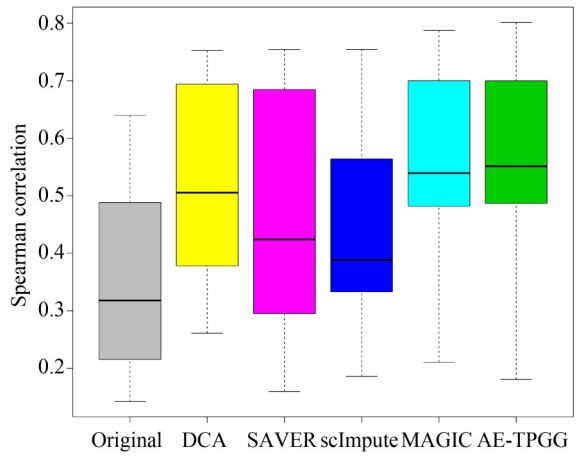



**Fig. 13** Spearman correlation coefficients of the six protein-RNA pairs for the original and imputed data using DCA, SAVER, scImpute, MAGIC and AE-TPGG

recovery ability of various methods, as shown in Fig. 15. It can be found that the correlation coefficients of the synthesized data with dropout present low values, indicating that the expression pattern of genes is obscured by dropout events. Although all imputation methods can ameliorate the recovery of missing pattern masked by dropout events, our model outperforms the other four methods and can restore the most accurate expression pattern of genes related to the development.

In addition, we especially select tbx-36 and his-8 genes with antagonistic expression patterns during the development of C.elegans [39] to examine the specific ability of AE-TPGG for recovering the expression patterns of individual genes. Fig. 16 displays gene expression trajectory for exemplary anti-correlated gene pair tbx-36 and his-8 over time for imputed results from AE-TPGG and the four baseline methods. From the figure it can be seen that synthesized data almost conceals the negative correlation of the two genes, but the imputed expression of the two genes obtained from AE-TPGG significantly restores the potential pattern relationship. In contrast, MAGIC and DCA also achieve some improvements on recovery of the antagonistic expression compared with AE-TPGG, while SAVER obtains an opposite relationship indicating the disability in pattern recovery for this pair of genes.

## 6 Discussion

scRNA-seq technologies have brought unprecedented capabilities to high-throughout, high-resolution transcriptomic analysis of cell states. Our method, AE-TPGG, is a joint learning paradigm for scRNA-seq data imputation and dimensionality reduction, which has the following highlights.

Firstly, the adopted TPGG distribution is based on the normalized scRNA-seq data. It is well known that data normalization is one of the most crucial steps of scRNA-seq data processing. Various experimental results also fully indicate that data normalization has a profound impact on the analysis of results. For example, the comparative experiments

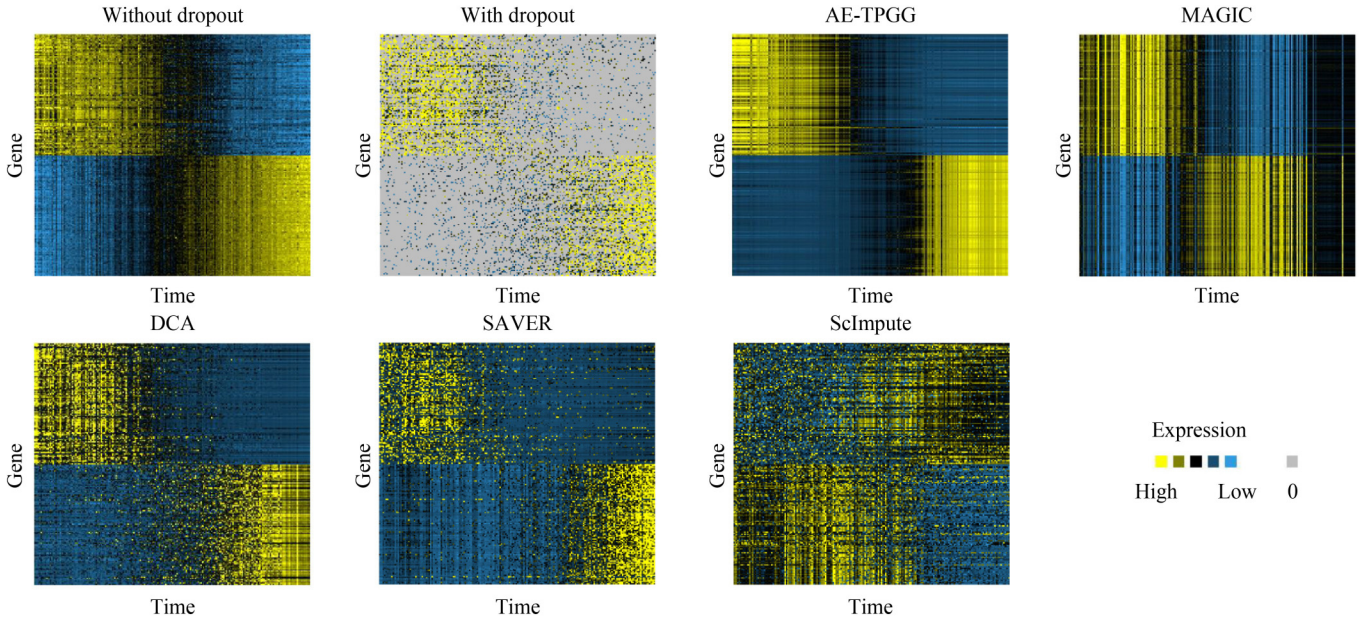



**Fig. 14** Heatmaps of the top 200 highly variable genes that consists of 100 positive genes and 100 negative genes associated with time course within the dataset using expression data without dropout, with dropout, and the imputed data from AE-TPGG, MAGIC, DCA, SAVER, scImpute, respectively. Yellow and blue colors represent relative high and low expression levels, respectively. Zero values are colored grey

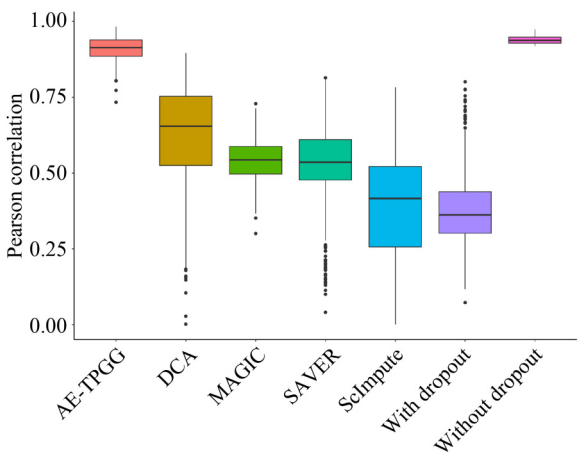



**Fig. 15** Boxplots of Pearson correlation coefficients between gene expression and the known developmental pattern across the 500 most highly correlated genes within the dataset using the imputed data from AE-TPGG, DCA, MAGIC, SAVER, scImpute, and expression data with dropout, without dropout, respectively. The box represents the interquartile range, the horizontal line in the box is the median, and the whiskers represent 1.5 times the interquartile range. Black dots represent outliers

between With dropout-C and With dropout-N in Section 5.1 indicate that data normalization conduces to improve the discovery of cell heterogeneity on the four datasets from visualization and clustering, even without imputation. Similarly, it is also observed that the imputed result obtained by MAGIC-N is obviously superior to MAGIC-C in clustering performance. The adopted TPGG distribution of AE-TPGG is based on the semi-continuous data, which also take into account data normalization. Therefore, modeling on normalized data is one of the reasons for our method to achieve competitive advantage. On the other hand, the difference between our method and DCA lies in the selection of distribution. DCA adopts ZINB distribution that aims at count data without normalization, while TPGG distribution is modeled on normalized data. As a result, our method outperforms DCA in various types of imputation experiments, which further illustrates the importance of data normalization for scRNA-seq data imputation.

Secondly, compared with other methods, our model takes better account of the distribution characteristics of normalized data, mainly including two points. One is a large number of dropout events, and the positive expression presents a typical right-skewed characteristic. In Section 3, the detailed statistical analysis in real scRNA-seq data is given to confirm the rationality of the two alternative models of TPGM and TPLNM. Furthermore, the adopted TPGG can adaptively adjust the parameters to select the appropriate right-skewed distribution for modeling the positive expression of genes, which is not explored by other imputation methods. This advantage can be easily observed from the comparative experiments of MAGIC and AE-TPGG. The imputation approach of MAGIC shares information across similar cells by means of graph diffusion, without considering the particular statistical characteristics of scRNA-seq data. In the experiment of Section 5.1, although the input data forms of the two methods were the same, clustering results indicate that AE-TPGG has better performance than MAGIC-N in scRNA-seq data imputation on four real datasets, especially in Deng dataset and Klein dataset.

Thirdly, the internal connection of the network of AE-TPGG automatically captures the correlation between genes, which contributes to promoting the imputation performance.

Finally, over-imputation and under-imputation are common issues in data imputation. The key to discuss these problems lies in the existence of benchmark data. Since the truth of scRNA-seq data is normally unknown, indirect assessment methods are frequently used by numerous imputation methods to validate the effect of scRNA-seq data imputation to a

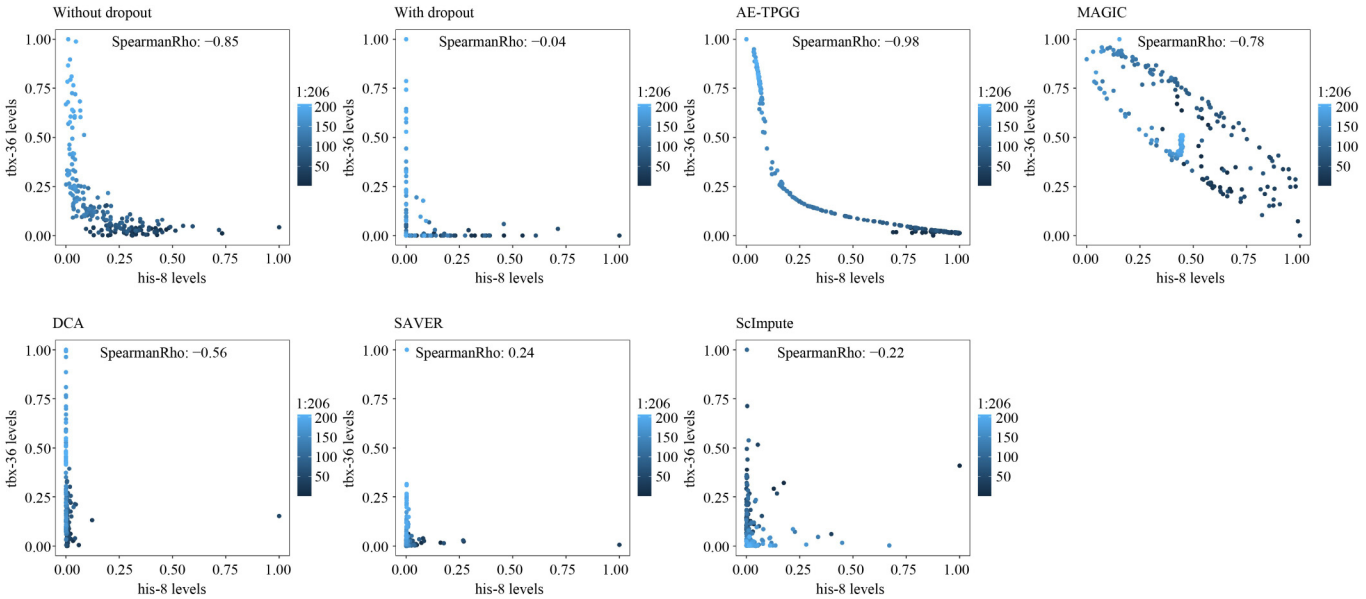



**Fig. 16** Gene expression trajectory for exemplary anti-correlated gene pair tbx-36 and his-8 over time for the data without, with dropout, and imputation data using AE-TPGG, MAGIC, DCA, SAVER and scImpute

certain extent. Specifically, in the experiments of Section 5.3, the cell surface marker protein expressions were taken as reference. Taking the enrichment of protein abundance at the RNA level as the benchmark, all imputation methods have the problems of over-imputation and under-imputation, and there are also differences between methods from the observations of Fig. 11, Fig. 12, and Fig. A1 in Appendix A, which means that no imputation method achieves the perfect match between the imputation data and the benchmark. Despite the different model assumptions among the imputation methods, the Spearman correlation coefficients displayed in Fig. 13 indicate that the imputation of scRNA-seq data of all methods enhances the correlation level of protein-RNA pairs. And especially, our method obtains better performance than other imputation methods. In Section 5.4 the microarray data were used as "ground truth" since the bulk transcriptomics contain less noise than single-cell transcriptomics. It can be observed that various imputation methods also have the issues of over-imputation and under-imputation from Fig. 14, but DCA, SAVER and our method have high consistency with the microarray data on the whole. In addition, Fig. 15 and Fig. 16 indicate that our method achieves substantially improved performance for recovering time-course patterns. Therefore, imputation deviations are hard to avoid, but our method can provide the most consistent results with the given benchmarks, indicating that a good imputation method can assist the analysis of scRNA-seq data.

## 7 Conclusion

In this work, we focus on the problem of frequent dropout events in scRNA-seq data, which hinders the downstream analyses. In response to this issue, we have proposed the AE-TPGG model, a deep autoencoder based on a Two-Part-Generalized-Gamma distribution, for the analysis of normalized scRNA-seq data. The proposed method considers characteristics of the semi-continuous normalized scRNA-seq data, and adopts the negative log-likelihood of TPGG as the loss to estimate the model parameters revealing the hidden low-dimensional structures and the inherent gene relationships. The proposed method achieves significant improvement for downstream analysis of scRNA-seq data in both dimensionality reduction and imputation. For the four real datasets, the imputed data with our model accurately captures the potential structures and the clustering results confirm the effectiveness of our method compared with the other four mainstream imputation methods. On the PMBC 68K dataset, we use the bottleneck layer to visualize the raw data in two-dimensional space so that the hidden structures of the data can present cell subtypes intuitively. Meanwhile, clustering results prove effective low-dimensional representation obtained from AE-TPGG compared with the original t-SNE method. Besides, the imputation data from our method on CBMCs improves protein and RNA co-expression analysis. In addition, we perform comprehensive studies on a semi-real dataset to compare AE-TPGG with other imputation approaches in recovering gene expression patterns with time course. Our work shows that AE-TPGG has the potential to improve the discovery of potential biological patterns from scRNA-seq data.

**Acknowledgements** This research was supported by the National Natural Science Foundation of China (Grant Nos. 62136004, 61802193), the National Key R&D Program of China (2018YFC2001600, 2018YFC2001602), the Natural Science Foundation of Jiangsu Province (BK20170934), and the Fundamental Research Funds for the Central Universities (NJ2020023). Thanks to all the open-minded researchers providing the codes and research resources. Thanks to all the anonymous reviewers.

## Appendix A

Fig. A1 The t-SNE visualizations of the protein expression (1st row), RNA expression derived from the original data (2nd row), imputation data using AE-TPGG (3rd row), MAGIC (4th row), DCA (5th row), SAVER (6th row) and scImpute

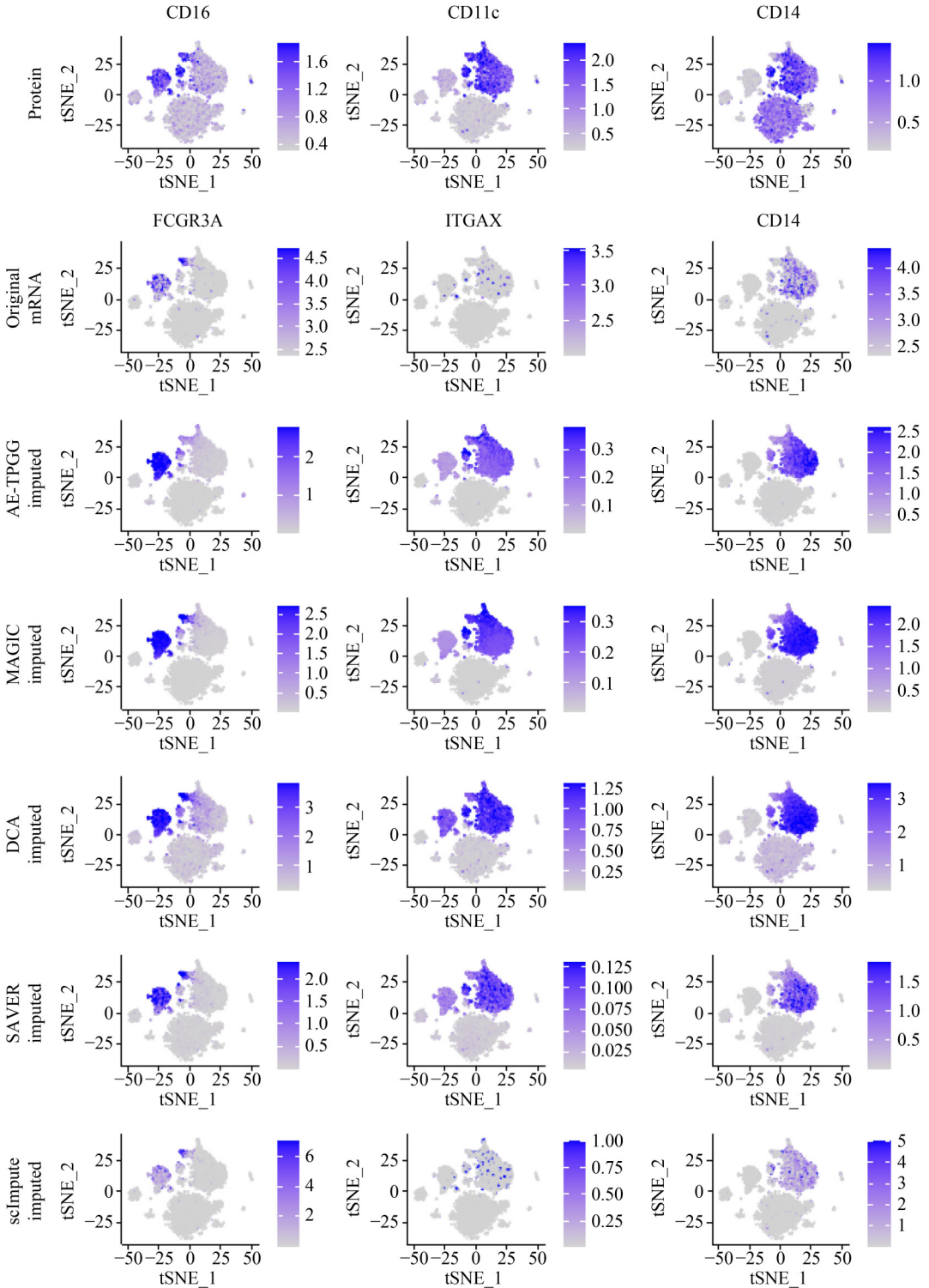



**Fig. A1** The t-SNE visualizations of the protein expression (1st row), RNA expression derived from the original data (2nd row), imputation data using AE-TPGG (3rd row), MAGIC (4th row), DCA (5th row), SAVER (6th row) and scImpute (7th row). Columns correspond to CD16 (1st column), CD11c (2nd column), CD14 (3rd column) proteins and corresponding RNAs FCGR3A, ITGAX and CD14

(7th row). Columns correspond to CD16 (1st column), CD11c (2nd column), CD14 (3rd column) proteins and corresponding RNAs FCGR3A, ITGAX and CD14.

## References

1. Potter S S. Single-cell RNA sequencing for the study of development, physiology and disease. Nature Reviews Nephrology, 2018, 14(8): 479–492

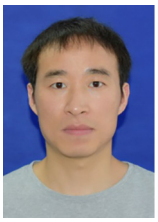
2. Li H, Courtois E T, Sengupta D, Tan Y, Chen K H, Goh J J L, Kong S L, Chua C, Hon L K, Tan W S, Wong M, Choi P J, Wee L J K, Hillmer A M, Tan I B, Robson P, Prabhakar S. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human

colorectal tumors. Nature Genetics, 2017, 49(5): 708–718

3. Cao Y, Su B, Guo X, Sun W, Deng Y, Bao L, Zhu Q, Zhang X, Zheng Y, Geng C, Chai X, He R, Li X, Lv Q, Zhu H, Deng W, Xu Y, Wang Y, Qiao L, Tan Y, Song L, Wang G, Du X, Gao N, Liu J, Xiao J, Su X, Du Z, Feng Y, Qin C, Qin C, Jin R, Xie X S. Potent neutralizing antibodies against SARS-CoV-2 identified by high-throughput single-cell sequencing of convalescent patients' B cells. Cell, 2020, 182(1): 73–84.e16
4. Kharchenko P V, Silberstein L, Scadden D T. Bayesian approach to single-cell differential expression analysis. Nature Methods, 2014, 11(7): 740–742
5. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek A K, Slichter C K, Miller H W, Mcelrath M J, Prlic M, Linsley P S, Gottardo R. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biology, 2015, 16(1): 278
6. Lun A T L, Bach K, Marioni J C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biology, 2016, 17(1): 75
7. Li W V, Li J J. An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nature Communications, 2018, 9(1): 997
8. Huang M, Wang J, Torre E, Dueck H, Shaffer S, Bonasio R, Murray J I, Raj A, Li M, Zhang N R. SAVER: gene expression recovery for single-cell RNA sequencing. Nature Methods, 2018, 15(7): 539–542
9. Van Dijk V, Sharma R, Nainys J, Yim K, Kathail P, Carr A J, Burdziak C, Moon K R, Chaffer C L, Pattabiraman D, Bierie B, Mazutis L, Wolf G, Krishnaswamy S, Pe'er D. Recovering gene interactions from single-cell data using data diffusion. Cell, 2018, 174(3): 716–729.e27
10. Basharat Z, Majeed S, Saleem H, Khan I A, Yasmin A. An overview of algorithms and associated applications for single cell RNA-seq data imputation. Current Genomics, 2021, 22(5): 319–327
11. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436–444
12. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798–1828
13. Hornik K. Approximation capabilities of multilayer feedforward networks. Neural Networks, 1991, 4(2): 251–257
14. Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. Science, 2006, 313(5786): 504–507
15. Kadurin A, Nikolenko S, Khrabrov K, Aliper A, Zhavoronkov A. druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. Molecular Pharmaceutics, 2017, 14(9): 3098–3104
16. Eraslan G, Simon L M, Mircea M, Mueller N S, Theis F J. Single-cell RNA-seq denoising using a deep count autoencoder. Nature Communications, 2019, 10(1): 390
17. Zhang Z, Cui F, Wang C, Zhao L, Zou Q. Goals and approaches for each processing step for single-cell RNA sequencing data. Briefings in Bioinformatics, 2021, 22(4): bbaa314
18. Mortazavi A, Williams B A, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-seq. Nature Methods, 2008, 5(7): 621–628
19. Pickrell J K, Marioni J C, Pai A A, Degner J F, Engelhardt B E, Nkadori E, Veyrieras J B, Stephens M, Gilad Y, Pritchard J K. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature, 2010, 464(7289): 768–772
20. Risso D, Schwartz K, Sherlock G, Dudoit S. GC-content normalization for RNA-seq data. BMC Bioinformatics, 2011, 12(1): 480
21. Vallejos C A, Risso D, Scialdone A, Dudoit S, Marioni J C. Normalizing single-cell RNA sequencing data: challenges and opportunities. Nature Methods, 2017, 14(6): 565–571
22. Li B, Ruotti V, Stewart R M, Thomson J A, Dewey C N. RNA-seq gene expression estimation with read mapping uncertainty. Bioinformatics, 2010, 26(4): 493–500
23. Belotti F, Deb P, Manning W G, Norton E C. Twopm: two-part models. The Stata Journal: Promoting communications on statistics and Stata, 2015, 15(1): 3–20
24. Lawless J F. Inference in the generalized gamma and log gamma distributions. Technometrics, 1980, 22(3): 409–419
25. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert J P. A general and flexible method for signal extraction from single-cell RNA-seq data. Nature Communications, 2018, 9(1): 284
26. Klein A M, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz D A, Kirschner M W. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell, 2015, 161(5): 1187–1201
27. Minka T P. Estimating a gamma distribution. Microsoft Research, 2002, 1(3): 3–5
28. Chollet F. Keras. See Github.com/fchollet/keras website
29. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems. 2016, arXiv preprint arXiv: 1603.04467
30. Deng Q, Ramsköld D, Reinius B, Sandberg R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science, 2014, 343(6167): 193–196
31. Kolodziejczyk A A, Kim J K, Tsang J C H, Ilicic T, Henriksson J, Natarajan K N, Tuck A C, Gao X, Bühler M, Liu P, Marioni J C, Teichmann S A. Single cell RNA-Sequencing of pluripotent states unlocks modular transcriptional variation. Cell Stem Cell, 2015, 17(4): 471–485
32. Ding J, Adiconis X, Simmons S K, Kowalczyk M S, Hession C C, Marjanovic N D, Hughes T K, Wadsworth M H, Burks T, Nguyen L T, Kwon J Y H, Barak B, Ge W, Kedaigle A J, Carroll S, Li S, Hacohen N, Rozenblatt-Rosen O, Shalek A K, Villani A C, Regev A, Levin J Z. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. Nature Biotechnology, 2020, 38(6): 737–746
33. Arthur D, Vassilvitskii S. k-means++: the advantages of careful seeding. In: Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms. 2007, 1027–1035
34. Zheng G X Y, Terry J M, Belgrader P, Ryvkin P, Bent Z W, Wilson R, Ziraldo S B, Wheeler T D, McDermott G P, Zhu J, Gregory M T, Shuga J, Montesclaros L, Underwood J G, Masquelier D A, Nishimura S Y, Schnall-Levin M, Wyatt P W, Hindson C M, Bharadwaj R, Wong A, Ness K D, Beppu L W, Deeg H J, Mcfarland C, Loeb K R, Valente W J, Ericson N G, Stevens E A, Radich J P, Mikkelsen T S, Hindson B J, Bielas J H. Massively parallel digital transcriptional profiling of single cells. Nature Communications, 2017, 8(1): 14049
35. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay P K, Swerdlow H, Satija R, Smibert P. Simultaneous epitope and transcriptome measurement in single cells. Nature Methods, 2017, 14(9): 865–868
36. Xu C, Su Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. Bioinformatics, 2015, 31(12): 1974–1980
37. Levine J, Simonds E, Bendall S, Davis K, Amir E A, Tadmor M, Litvin O, Fienberg H, Jager A, Zunder E, Finck R, Gedman A, Radtke I,

Downing J, Pe'er D, Nolan G. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. Cell, 2015, 162(1): 184–197

38. Francesconi M, Lehner B. The effects of genetic variation on gene expression dynamics during development. Nature, 2014, 505(7482): 208–211
39. Boeck M E, Huynh C, Gevirtzman L, Thompson O A, Wang G, Kasper D M, Reinke V, Hillier L W, Waterston R H. The time-resolved transcriptome of *C. elegans*. Genome Research, 2016, 26(10): 1441–1450



Shuchang Zhao received the BSc degree from the Suzhou University, China in 2013, and MSc degree from the Anhui University of Science and Technology, China in 2016, respectively. Currently, he is working toward the PhD degree in the PARNEC group of the College of Computer Science and Technology at Nanjing University of Aeronautics and Astronautics (NUAA), China. His research interests include machine learning and bioinformatics.

Li Zhang received the BSc degree from the Changsha University of Science and Technology, China in 2007, and the MSc and PhD degrees from the Nanjing University of Aeronautics and Astronautics (NUAA), China in 2010 and 2015, respectively. He joined the College of Computer Science and Technology, Nanjing Forestry University, as a Lecturer, China in 2016. His current research interests include machine learning and bioinformatics.

Xuejun Liu received the BSc and MSc degrees in computer science from the Nanjing University of Aeronautics and Astronautics (NUAA), China in 1999 and 2002, respectively, and the PhD degree in computer science from the University of Manchester, UK in 2006. Currently, she is a professor in the PARNEC group of the College of Computer Science and Technology at NUAA, China. Her research interests include machine learning and its practical applications, including bioinformatics.