Multiobjective Deep Clustering and Its Applications in Single-cell RNA-seq Data

Yunhe Wang, Chuang Bian, Ka-Chun Wong, Xiangtao Li, and Shengxiang Yang

***Abstract*—Single-cell RNA sequencing is a transformative technology that enables us to study the heterogeneity of the tissue at the cellular level. Clustering is used as the key computational approach to group cells under the transcriptome profiles from single-cell RNA-seq data. However, accurate identification of distinct cell types is facing the challenge of high-dimensionality, and it could cause uninformative clusters when clustering is directly applied on the original transcriptome. To address such challenge, an evolutionary multiobjective deep clustering algorithm (EMDC) is proposed to identify single-cell RNA-seq data in this study. First, EMDC removes redundant and irrelevant genes by applying the differential gene expression analysis to identify differentially expressed genes across biological conditions. After that, deep autoencoder is proposed to project the high-dimensional data into different low-dimensional nonlinear embedding subspaces under different bottleneck layers. Then, the basic clustering algorithm is applied in those nonlinear embedding subspaces to generate some basic clustering results to produce the cluster ensemble. To lessen the unnecessary cost produced by those clusterings in the ensemble, the multiobjective evolutionary optimization is designed to prune the basic clustering results in the ensemble, unleashing its cell type discovery performance under three objective functions. Multiple experiments have been conducted on thirty synthetic single-cell RNA-seq datasets and six real single-cell RNA-seq datasets, which reveal that EMDC outperforms eight other clustering methods and three multiobjective optimization algorithms in cell type identification. In addition, we have also conducted extensive comparisons to effectively demonstrate the impact of each component in our proposed EMDC.**

***Index Terms*—Single-cell RNA-seq dataset, Evolutionary Multiobjective Deep Clustering, Multiobjective Optimization.**

I. Introduction

SINGLE-CELL RNA sequencing technologies have emerged to reveal the cell expression across the whole genome at single cell resolution [1]. It employs single cells as sequencing samples to address fundamental questions that cannot be solved by bulk-level experiments, which can reveal unappreciated levels of heterogeneity, describe RNA molecules in individual cells with high resolution, and characterize tumor microevolution [2]. Therefore, single-cell RNA sequencing can be widely used to understand complex biological systems. Accurate identification of distinct cell types from numerous heterogeneous cells is an indispensable task in single-cell RNA-seq data analysis. To provide an intuitive way for this task, clustering is the key computational technique since it can separate cells into different clusters for characterizing cell subtypes based on their transcriptome profiles [3]. Moreover, the clustering results can provide the diverse downstream analysis of single-cell RNA-seq data [4]. Multiple kinds of clustering algorithms have been developed to group different cells into distinct cell types; for instance, K-means algorithm [5], hierarchical clustering such as CIDR [6], community detection methods [7], and other similarity learning approaches like SIMLR [8]. However, it is hard to believe that a single unsupervised clustering method can achieve the all-time best clustering results for various single-cell RNA-seq data. It is rather challenging to choose a suitable clustering algorithm for the single-cell RNA-seq data since distinct clustering methods perform differently on those datasets with their unique advantages and disadvantages.

Ensemble learning can integrate the clustering results obtained from several base clustering algorithms into a consensus clustering. It has been widely applied in a variety of fields [9]. For example, Asur *et al.* [10] proposed an ensemble clustering framework to group the protein-protein interaction networks. Huang *et al.* [11] devised a locally weighted ensemble clustering method on real-world datasets based on the local weighting strategy. Yang *et al.* [12] developed a single-cell aggregated clustering gathering multiple base clustering results to group single-cell RNA-seq data. However, these ensemble methods often suffer from the high level of technical noise, intrinsic biological variability, and high-dimensionality of the single-cell RNA-seq data. Intuitively, since these clustering algorithms overly rely on specific similarity metrics in the ensemble models, and similarity metrics between cells become meaningless in high-dimensional spaces, it may result in low-quality ensemble members.

A direct and practical approach is to embed the high-dimensional data into the low-dimensional latent space to capture the underlying structure from the original data. Different dimension reduction methods have been developed to extract the low-dimensional feature representations from the single-cell RNA-seq data. For example, Shin *et al.* [13]

Y.H. Wang is with the School of Artificial Intelligence, Hebei University of Technology, Tianjin, China and with the School of Computer Science and Informatics, De Montfort University, Leicester, UK.

C. Bian is with the School of Artificial Intelligence, Jilin University, Changchun, China.

K.C. Wong is with the Department of Computer Science, City University of Hong Kong, Hong Kong. E-mail: kc.w@cityu.edu.hk.

X.T. Li is with the School of Artificial Intelligence, Jilin University, Changchun, China. (Corresponding author: lixt314@jlu.edu.cn)

S. Yang is with the School of Computer Science and Informatics, De Montfort University, Leicester, UK. (Corresponding author: syang@dmu.ac.uk)

adopted the principal component analysis (PCA) [14], to project those high-dimensional single-cell RNA-seq data into lower-dimensional spaces using a linear transformation of the original variables with the largest variances. Grün *et al.* [15] employed the t-distributed stochastic neighbor embedding algorithm (t-SNE) [16] to visualize the single-cell RNA-seq data in two dimensions and then employed a K-means clustering algorithm. However, these methods have their own disadvantages: simple linear methods such as PCA cannot capture the non-linear characteristics of the gene expression data, while the non-linear methods such as t-SNE are very sensitive to the hyper-parameter setting and cannot learn a parametric mapping. To address those problems, Li *et al.* [9] proposed an evolutionary multiobjective ensemble pruning algorithm (EMEP) to identify cell types from the single-cell RNA-seq data by employing the non-negative matrix factorization (NMF) for dimension reduction. Unfortunately, NMF cannot guarantee the convergence. Moreover, it cannot balance well between data sparsity that represents the latent local feature and data interpretability [17].

Recently, deep autoencoder models have been developed as the dimension reduction method for identifying high-dimensional single-cell RNA-seq data, which can learn the potential manifold structure and capture non-linear complex dependencies under low-dimensional spaces from high-dimensional single-cell RNA-seq data. For instance, Wang and Gu [18] proposed a deep variational autoencoder for the single-cell RNA-seq datasets to discover the nonlinear hierarchical feature representations. Tangherloni *et. al* [19] developed a unifying tool based on the autoencoders to facilitate analyzing the single-cell RNA-seq data. Chen *et al.* [20] designed an adaptive fuzzy K-means algorithm combined with the deep autoencoder technique.

In this study, we propose an evolutionary multiobjective deep clustering algorithm (EMDC), which embeds the learning representation by the deep autoencoder into the evolutionary multiobjective clustering to identify the single-cell RNA-seq data from transcriptome profiles. First, EMDC applies the differential gene expression analysis to preselect numerous genes from the original high-dimensional single-cell RNA-seq data. Second, deep autoencoder is employed to map those data into different low-dimensional latent subspaces. After that, we apply a basic clustering algorithm on those learned low-dimensional latent spaces to generate the cluster ensemble. Then, multiobjective evolutionary optimization is designed to prune the basic clustering results in the ensemble to further enhance the generalization performance under three objective functions. To guide the evolution, three objective functions including Davies-Bouldin Index (Db), Dunn Validity Index (Dunn), and the number of clusterings in the ensemble are proposed, where Db and Dunn are two internal cluster validity indices for evaluating the clustering performance. In the experiment, we apply our proposed EMDC and other comparative methods on thirty synthetic single-cell RNA-seq datasets and six real single-cell RNA-seq datasets to reveal the effectiveness of our proposed EMDC. Other extended analyses are also performed to demonstrate the robustness of each component in EMDC from different perspectives.

II. Methods

A. Methodology Overview of EMDC

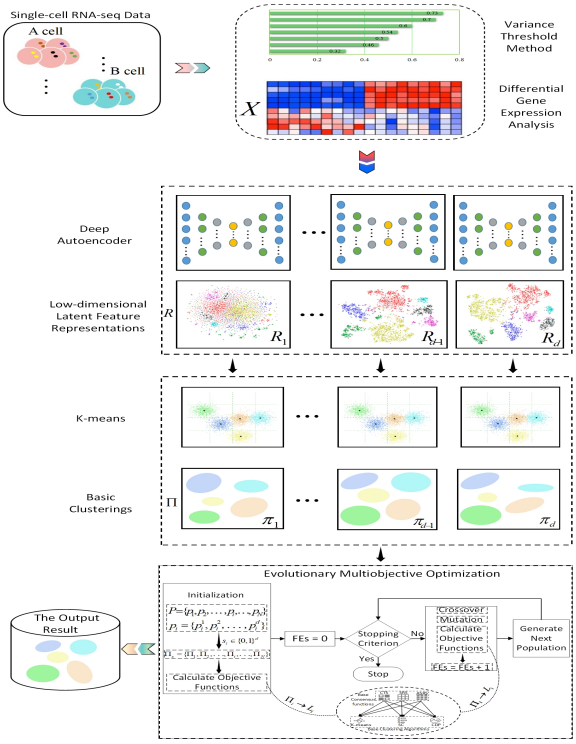


Fig. 1: The overview of EMDC. In the first, the differential gene expression analysis and variance threshold method reduction are adopted to generate the gene expression matrix X. After that, the deep autoencoder is proposed to project X into different low-dimensional latent feature representations $R = \{R_1, R_2, ..., R_d\}$. Then, we employed the basic clustering algorithm such as K-means clustering algorithm to produce the basic partitions $\Pi = \{\pi_1, \pi_2, ..., \pi_d\}$. Finally, an evolutionary multiobjective optimization is proposed to cluster the specific single-cell RNA-seq data under three objective functions including Db, Dunn, and the number of basic clusterings in the ensemble. It is worth noting that different consensus functions and base clustering algorithms can be selected adaptively throughout the evolution process.

In this study, EMDC is proposed for identifying single-cell RNA-seq data. We summarize the overview of EMDC in Fig. 1, which indicates that four components are encapsulated in EMDC. In the first component, we adopt the differential gene expression analysis [21] and variance threshold method [22] to preselect the top 2000 genes, generating the gene expression matrix X from the original single-cell RNA-seq data. After that, EMDC employs a deep autoencoder to project the selected gene expression matrix X into different latent feature representations $R = \{R_1, R_2, ..., R_j, ..., R_d\}$, the size of R_j ($j = \{1, 2, ..., d\}$) is $n \times m$, in which n is the number of samples, m denotes the dimension of latent feature representations ranging from $\{20, 30, 40, ..., 200\}$, d is the number of different low-dimensional latent feature representations, and the number of elements in the m value collection is equal to d. For example, given $m = 40$, it corresponds to $j = 3$ and R_3 denotes the third latent feature representation with the dimension of 40. Following that, the basic clustering algorithm is applied in those non-linear embedding subspaces to generate multiple basic clustering results to produce the cluster ensemble. The ensemble composed of d basic clusterings $\Pi = \{\pi_1, \pi_2, ..., \pi_d\}$ is produced.

In the last component, an evolutionary multiobjective optimization using Π is designed to interpret the single-cell RNA-seq data. Regarding that, we encode each individual in the population to produce different ensembles $\Pi_s = \{\Pi_1, \Pi_2, ..., \Pi_i, ..., \Pi_N\}$, in which N is the number of individuals in the population, Π_i ($i = \{1, 2, ..., N\}$) represents the i-th ensemble with a number of basic clusterings chosen from Π. To guide the evolution and capture different characteristics of the single-cell RNA-seq data, three objective functions including Db, Dunn, and the number of basic clusterings, are proposed and formulated. We map the ensemble with a number of basic clusterings for each individual in the population to calculate those objective functions through $\Pi_i \rightarrow L_i$ ($i = \{1, 2, ..., N\}$), where Π_i is the ensemble and L_i is a clustering result. It is worth noting that any of those base consensus functions and base clustering algorithms can be selected to promote the clustering performance during the evolutionary process. Each ensemble Π_i ($i = \{1, 2, ..., N\}$) has one binary mask vector $s_i = \left\{s_i^1, s_i^2, ..., s_i^j, ..., s_i^d\right\} \in \{0, 1\}^d$ to select the basic clusterings, in which $s_i^j = 1 (j = 1, 2, ..., d)$ indicates that the j-th basic clustering in the ensemble is chosen, otherwise it means that the j-th basic clustering is not selected.

B. Differential Gene Expression Analysis

In EMDC, differential gene expression analysis is adopted to select some differentially expressed genes from the single-cell RNA-seq data. For differentially expressed genes detection under linear model, we employ a differential gene expression software package limma [21] which is freely available.

First, the single-cell RNA-seq data is transformed to logarithmic scale and then we sequentially compare the samples in one group with all the samples of the other groups, calculating the significant difference and the fold change. The significant difference is computed by the hypothesis testing and measured by the p-value. Considering g_1 and g_2 as two genes of the single-cell RNA-seq data, we assume that there is no difference between the expressions of g_1 and g_2. Based on this null hypothesis, the p-value is obtained by T-test. If the p-value is less than 0.05, a small probability event has occurred and the null hypothesis is rejected. Therefore, the expressions of g_1 and g_2 are significantly different. However, we discover a phenomenon called overdispersion on those data, which makes Poisson-based analysis prone to high false positive rates. Since the differentially expressed genes detection involves a large number of statistical tests, we need to take the multiplicity into account to determine the detection significantly. Therefore, in our study, we control the expected rate of false positives in all the detections and correct the p-values by the Benjamini-Hochberg method [23]. It adopts the q-value as the key indicator to screen differentially expressed genes. A gene is a differentially expressed gene when its q-value is less than 0.05. For calculating the q-values, all those p-values are sorted in the ascending order, which can be defined as follows:

$$q_i = \frac{p_i \times m_p}{i} \tag{1}$$

where m_p is the total number of p-values, $p_i (i = \{1, 2, ..., m_p\})$ denotes the i-th p-value after sorting, and q_i is the i-th q-value. Finally, the fold change can be calculated as follows:

$$FC = \frac{avg(t_1)}{avg(t_2)} \tag{2}$$

where t_1 and t_2 are the expression values of the genes in the t_1-th and t_2-th group respectively, and $avg(\cdot)$ is the average value function. Each gene has the fold change of the expression values in different groups. If the q-value of the gene is less than 0.05, then we calculate its fold change. The first 1000 genes sorted by the absolute value of $log_2(FC)$ in the descending order will be taken. Moreover, the variance threshold method is employed to preselect another 1000 genes from the high-dimensional single-cell RNA-seq data. Specifically, we remove the duplicated genes to generate the final genes for EMDC.

C. Deep Autoencoder

After differential gene expression analysis, deep autoencoder is proposed in EMDC to generate different latent feature representations with low dimension. Deep autoencoder is a multi-layer feedforward neural network that consists of an encoder and a decoder, intersecting at the bottleneck layer with lower dimension compared with the input dimension. It aims to reconstruct the high-dimensional data with the minimum error, learning low-dimensional latent feature representations for the output data of the bottleneck layer. In our study, d latent feature representations under different bottleneck layers $R = \{R_1, R_2, ..., R_j, ..., R_d\}$ ($j = \{1, 2, ..., d\}$) are constructed, in which $R_j = \left\{r_{j(1)}, r_{j(2)}, ..., r_{j(i)}, ..., r_{j(n)}\right\}$, n is the number of cell samples, $r_{j(i)} = \left\{r_{j(i)}^1, r_{j(i)}^2, ..., r_{j(i)}^k, ..., r_{j(i)}^m\right\}$ ($i = \{1, 2, ..., n\}, k = \{1, 2, ..., m\}, m = \{20, 30, 40, ..., 200\}$), $r_{j(i)}^k$ is the k-th dimension of the i-th sample in the representation R_j. First, we train deep autoencoder by the preselected dataset $X = \{x_1, x_2, ..., x_i, ..., x_n\}$ with n cells and each cell x_i contains $\tilde{m}$ genes which can be represented as $x_i = \left\{x_i^1, x_i^2, ..., x_i^{\tilde{m}}\right\}$. Fig. 2 shows the schematic view of deep autoencoder in our proposed EMDC. It includes a flattened input layer that denotes the gene expression profiles $X = \{x_1, x_2, ..., x_n\}$, several hidden layers, one bottleneck layer which is denoted as $R_j = \left\{r_{j(1)}, r_{j(2)}, ..., r_{j(i)}, ..., r_{j(n)}\right\}$, and an output layer that denotes the reconstructed gene expression profiles $\widehat{X} = \{\widehat{x}_1, \widehat{x}_2, ..., \widehat{x}_n\}$. The encoder contains the input layer, the bottleneck layer, and two hidden layers while the decoder includes the bottleneck layer, two hidden layers, and the output layer. All those hidden layers own different hidden units. Besides, the neural unit connection between layers is fully connected.

In deep autoencoder of EMDC, the encoder ($\varphi_\theta(\cdot)$) can compress and map the input vectors x_i to generate a low-dimensional representation vector $r_{j(i)}$ in the latent space R_j. Then, that representation vector is accepted as the input of the decoder. While the decoder is to map that representation $r_{j(i)}$ back to the space with high dimension and obtain a reconstructed vector $\widehat{x}_i$. The encoder and decoder that are composed of multiple neuron layers can be defined as follows:

$$\begin{aligned} \varphi_\theta^l(\cdot) &= \sigma^l(W^l(\varphi_\theta^{l-1}(\cdot)) + b^l) \\ \phi_\theta^l(\cdot) &= \sigma^l(\widehat{W}^l(\phi_\theta^{l-1}(\cdot)) + \widehat{b}^l) \end{aligned} \tag{3}$$

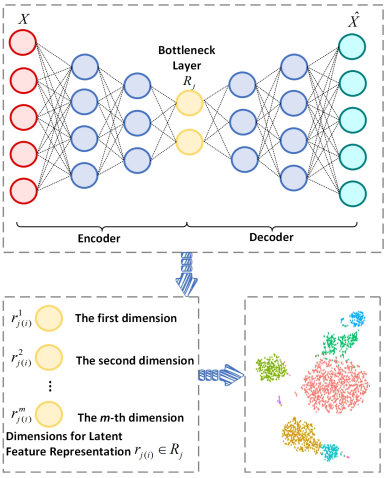


Fig. 2: The schematic view of deep encoder in EMDC.

where σ is the activation function, θ is the model parameter, l denotes the layer index, W and $\widehat{W}$ represent the weight matrices, b and $\widehat{b}$ represent the bias vectors.

In EMDC, Exponential Linear Unit (ELU) is used as the non-linear activation function in the layers, which is defined as below:

$$\sigma(x) = \begin{cases} x, \ if \ x > 0 \\ \alpha(e^x - 1), \ if \ x \leqslant 0 \end{cases} \tag{4}$$

where α denotes the non-zero gradient.

After that, deep autoencoder can train the model by optimizing the objective function as follows:

$$\begin{aligned} r_{j(i)} &= \varphi_\theta(x_i) \\ \widehat{x}_i &= \phi_\theta(r_{j(i)}) \\ \min \ Loss(x_i, \widehat{x}_i) &= \min(\frac{1}{\tilde{m}} \sum_{k=1}^{\tilde{m}} (x_i^k - \widehat{x}_i^k)^2) \end{aligned} \tag{5}$$

where x_i is the input samples, $r_{j(i)}$ is the low-dimensional feature representation vector of the input original single-cell RNA-seq data with high dimension under the bottleneck layer, $\widehat{x}_i$ is its reconstruction vector, $\tilde{m}$ is the number of genes in the input data, and $Loss(\cdot)$ is the loss function that can measure the differences between x_i and $\widehat{x}_i$. In addition, we initialize the weight matrix by the he_uniform method [24] and adopt the Adam algorithm [25] as the optimizer with a learning rate of 0.0001.

D. Objective Functions

After mapping the high-dimensional data into different low-dimensional latent feature representations, EMDC is proposed for identifying cell types by evolving several objective functions. To guide the evolution, suitable objective functions should be devised. It is worth noting that the goals of EMDC are to enhance its generalization ability and reduce the number of base clusterings in the selected basic partition. In order to achieve those goals, three objective functions including Db [26], Dunn [27], and Nc [9] are considered. The first and second objective functions are employed to enhance the generalisability of EMDC while the third objective function is utilized to reduce the number of clustering members.

The first objective function calculates the intra-cluster similarity of the clustering. Smaller values indicate better clustering results, which is defined as:

$$\begin{aligned} Db &= \frac{1}{K} \sum_{i=1}^{K} \max_{i \neq j} (\frac{\delta_i + \delta_j}{d(c_i, c_j)}) \\ \delta_i &= \sqrt{\frac{1}{n_k} \sum_{p \in C_i} \sum_{q=1}^{K} |x_{pq} - c_{iq}|^2} \end{aligned} \tag{6}$$

where K is the number of clusters in the predicted clustering, C_i represents the i-th cluster, c_i and c_j are two cluster centroids, and $d(c_i, c_j)$ is the Euclidean distance between c_i and c_j.

The second objective function is to find compact and well-separated clusters in the clustering. Larger values represent better clustering results. It can be defined as follows:

$$Dunn = \min_i \left\{ \min_j (\frac{\min_{x \in C_i, y \in C_j} d(x,y)}{\max_k \max_{x,y \in C_k} d(x,y)}) \right\} \tag{7}$$

where C_i, C_j, and C_k are the i-th, j-th, and k-th clusters in the predicted clustering, x, y are two data samples, and $d(x, y)$ denotes the Euclidean distance between x and y.

The third objective function is to minimize the number of base clusterings in the selected basic partitions. In EMDC, Π_i denotes the i-th ensemble which comprises a number of basic clusterings, its binary mask vector is $s_i = \left\{ s_i^1, s_i^2, ..., s_i^j, ..., s_i^d \right\} \in \{0, 1\}^d$ in which $s_i^j = 1 (j = 1, 2, ..., d)$ means that the basic clustering $\pi_j \in \Pi$ is chosen and otherwise π_j is not selected. Hence, the third objective function can be computed as follows:

$$Nc = \|s_i\| = \sum_{j=1}^{d} s_i^j \tag{8}$$

E. Evolutionary Multiobjective Optimization for EMDC

Following the above sections, an evolutionary multiobjective optimization method for EMDC is designed to interpret the single-cell RNA-seq data. It contains the initialization, crossover, mutation, and the Pareto optimal approach. The initialization is employed for encoding the individuals among the population. The crossover and mutation operators focus on updating the individuals. After that, the Pareto optimal method aims to evolve the population based on the hypervolume for the multiobjective single-cell RNA-seq clustering problems.

1) Initialization: A population $P = \{p_1, p_2, ..., p_N\}$ with N individuals is constructed at the beginning. Each individual $p_i = \left\{ p_i^1, p_i^2, ..., p_i^d \right\}, i = \{1, 2, ..., N\}$ is randomly generated as follows:

$$\begin{aligned} p_i^j &= p_{min}^j + rand(0,1) \times (p_{max}^j - p_{min}^j), j \in \{1, 2, ..., d\} \\ s_i^j &= \begin{cases} 1, \ if \ p_i^j \leqslant 0.5 \\ 0, \ if \ p_i^j > 0.5 \end{cases} \end{aligned} \tag{9}$$

where $p_{max} = \left\{ p_{max}^1, p_{max}^2, ..., p_{max}^d \right\}$ is the upper bound while $p_{min} = \left\{ p_{min}^1, p_{min}^2, ..., p_{min}^d \right\}$ is the lower bound. $s_i^j = 1$ denotes that the basic clustering $\pi_j \in \Pi$ is chosen and otherwise it means that the basic clustering π_j is removed for the i-th individual p_i. By this coding, each individual

can choose the basic clustering subsets as an ensemble Π_i ($i = \{1, 2, ..., N\}$).

2) Crossover and Mutation: Following the initialization phase, the mutation and crossover operators are employed for the evolution phase. To optimize those objective functions, the simulated binary (SBX) crossover and polynomial mutation are proposed as the crossover and mutation operators respectively [28, 29]. The SBX crossover operator is employed on parent individuals under the tournament selection to generate N offspring individuals, which can be calculated as follows:

$$\begin{aligned} v_k^j &= 0.5 \times [(1+\gamma_j)p_{i_1}^j + (1-\gamma_j)p_{i_2}^j] \\ v_i^j &= 0.5 \times [(1-\gamma_j)p_{i_1}^j + (1+\gamma_j)p_{i_2}^j] \\ \gamma_j &= \begin{cases} (2\mu_j)^{\frac{1}{\eta_c+1}}, \ if \ \mu_j \leqslant 0.5 \\ \frac{1}{2(1-\mu_j)}^{\frac{1}{\eta_c+1}}, \ if \ \mu_j > 0.5 \end{cases} \\ \gamma_j &= \gamma_j \times (-1)^{r_1} \\ \gamma_j &= \begin{cases} 1, \ if \ r_2 > CR \\ \gamma_j, \ otherwise \end{cases} \end{aligned} \tag{10}$$

where $i, k, i_1, i_2 \in \{1, 2, ..., N\}$, $j \in \{1, 2, ..., d\}$, p_{i_1} and p_{i_2} are two parent individuals, v_k and v_i are two offspring individuals generated by the crossover operator, η_c is the distribution index, μ_j and r_2 are random numbers generated from the range $[0, 1]$, r_1 is the number randomly chosen from $\{0, 1\}$, and *CR* is the crossover rate.

After that, the polynomial mutation operator is implemented by mutating an offspring individual to produce a new individual under the mutation probability (*MP*), which can be described as follows:

$$\begin{aligned} v_i^j &= v_i^j + \sigma_j \times (p_{max}^j - p_{min}^j) \\ \sigma_j &= \begin{cases} [2\mu_j + (1-2\mu_j)(1-\sigma_1)^{\eta_m+1}]^{\frac{1}{\eta_m+1}} - 1, \ if \ \mu_j < 0.5 \\ 1 - [2(1-\mu_j) + 2(\mu_j - 0.5)(1-\sigma_2)^{\eta_m+1}]^{\frac{1}{\eta_m+1}}, \ if \ \mu_j \geqslant 0.5 \end{cases} \\ \sigma_1 &= \frac{p_i^j - p_{min}^j}{p_{max}^j - p_{min}^j} \\ \sigma_2 &= \frac{p_{max}^j - p_i^j}{p_{max}^j - p_{min}^j} \\ s_i^j &= \begin{cases} 1, & if \quad v_i^j \leqslant 0.5 \\ 0, & if \quad v_i^j > 0.5 \end{cases} \end{aligned} \tag{11}$$

where $j \in \{1, 2, ..., d\}$, v_i is the i-th offspring individual to be mutated, η_m is the distribution index, and μ_j is a random number within the range $[0, 1]$. After obtaining the offspring individual v_i, we transfer it into s_i that belongs to the binary space, to select base clustering members from all those basic clusterings.

3) Pareto Optimal Approach: In EMDC, three objective functions, including Db, Dunn, and Nc, are designed to evolve the population. However, those objective functions are always conflicting with each other. Hence, we propose the Pareto optimal approach to analyze single-cell RNA-seq data clustering problem under those objective functions. The multiobjective optimization plays the role of optimizing more than one conflicting objective functions simultaneously, satisfying all the constraints of equality and inequality. It is characterized as follows:

$$\begin{aligned} \min \quad & F_i(p), \quad i = 1, 2, ..., M \\ subject\ to \quad & g_j(p) \leq 0, \quad j = 1, 2, ..., J \\ & h_k(p) = 0, \quad k = 1, 2, ..., K \end{aligned} \tag{12}$$

where $p = \left\{p^1, p^2, ..., p^d\right\}$ is a candidate solution with d variables, M denotes the number of objective functions, J and K are the numbers of inequality and equality constraints. Taking a minimization problem, given p_1 dominates p_2, if for all objective functions, $F(p_1)$ is less than or equal to $F(p_2)$, and for at least one objective function, $F_i(p_1)$ is less than $F_i(p_2)$. p^* is regarded as a Pareto-optimal solution (a non-dominated solution) if and only if there is no solution dominating p^*. Based on it, single-cell RNA-seq clustering problems can be transferred to search the Pareto set composed of optimal clustering results using the multiobjective optimization.

The fast hypervolume-based algorithm (HypE) is widely used for the multiobjective optimization problem involving at least three objective functions [30]. It employs the Monte Carlo simulation in the fitness assignment method to approximate the fitness values. This novel mechanism can trade off the accuracy of fitness and the overall computing time budget, exploiting the potential of the hypervolume calculation indicator effectively. Inspired by [30], we optimize those non-dominated solutions by integrating the fast hypervolume-based algorithm to apply the the Monte Carlo simulation in fitness assignment mechanism to the mating and environmental selection. Algorithm 1 presents the framework of the evolutionary multiobjective optimization method for EMDC.

Algorithm 1: Evolutionary Multiobjective Optimization for EMDC

Input:
(1) Population size (N);
(2) The crossover rate (*CR*) and the mutation probability (*MP*);
Output:
The Pareto set $\overline{PS}$;
Initialize the population $P = \{p_1, p_2, ..., p_N\}$;
Transfer P into the binary space $\{0, 1\}^d$ to generate different ensembles for P;
Each individual in P is randomly assigned a base consensus function and the associated base clustering algorithm;
Calculate those three objective functions F for each individual in P;
while *the stopping criterion is not met* **do**
  Generate a mating pool $\tilde{P}$ based on hypervolume-based estimation method from P;
  $V \leftarrow$ Crossover($\tilde{P}$,*CR*);
  $V \leftarrow$ Mutation(V,*MP*);
  Transfer V into the binary space $\{0, 1\}^d$;
  Generate ensembles for V;
  Calculate F for each individual in V;
  Select N individuals to construct a new population U from $P \cup V$ based on the hypervolume;
  $P \leftarrow U$;
Return the Pareto set $\overline{PS}$ with all the non-dominated solutions under those three objective functions;

Firstly, a population P consisting of N individuals is initialized in the continuous search space. Secondly, the population P is transferred into the binary space $P \in \{0, 1\}^d$. Each individual is encoded to produce an ensemble Π_i. We put the CTS matrix [31], the SRS matrix [32], and the ASRS matrix [33] into the base consensus function pool to exhibit their diverse characteristics in EMDC. While K-means clustering algorithm [34], Spectral Clustering Algorithm (SC) [35], and Clustering by Fast Search and Find of Density Peaks (CDP)

[36] are chosen to compose the base clustering algorithm pool to interpret the single-cell RNA-seq data comprehensively [9]. Then, each individual p_i is assigned with one base consensus function and the associated base clustering algorithm selected randomly from the base consensus function pool and base clustering algorithm pool. After that, we compute those three objective functions, which are Db, Dunn, and Nc, to assess the performance of each individual. Meanwhile, a mating pool is generated utilizing the hypervolume-based estimation method on the current population P. Following that, the crossover and mutation operators are executed to obtain an offspring population V. We transfer V into the binary space and calculate those three objective functions for each individual. After that, a new population U with N individuals is updated by the hypervolume-based multiobjective environment selection from the multiset-union $P \cup V$. It can be noted that the update process by creating the new population follows two rules. The first one is the nondominated sorting principle, which is the frequently-used scheme in the hypervolume-based multiobjective optimizers [37, 38]. The process of this rule is that the multiset-union is divided into disjoint partitions using it and the new population is filled with the partition starting from the lowest dominance depth level [30]. The other one is to remove the individual with the worst fitness from the partition in each step [30]. Indeed, we choose one individual randomly and uniformly if several individuals have the same worst fitness. The whole process is repeated until the remaining positions in the population are filled. Finally, the Pareto set that includes all the non-nominated solutions is achieved. Meanwhile, the time complexity of EMDC is discussed in Supplementary Section I.

According to the characteristics of the multiobjective optimization algorithms, it is difficult to discover a global optimal solution on each iteration for the proposed algorithm. To address this issue, the independent objective-based approach has been proposed to choose the solution with the best value for a cluster validity index (not the objective function of the algorithm) as the final clustering solution [39]. Therefore, in our study, we use the normalized mutual information to measure each individual in the population and then pick up the best value from the non-dominated solutions as the final clustering solution [40].

III. Experiments and Results

A. Data Collection

To evaluate the performance of EMDC, thirty synthetic single-cell RNA-seq datasets and six real single-cell RNA-seq datasets are collected. The thirty synthetic single-cell RNA-seq datasets are produced by SPsimSeq [41], a semi-parametric simulation procedure. It is designed to maximally retain the characteristics of real RNA-seq data for simulating a wide range of scenarios. Each dataset is divided into three clusters. The first ten datasets contain 100 cells with genes ranged from 2000 to 6000. For the other datasets, the number of cells is varied within the set $\{200, 300, 400, 500\}$ and the number of genes ranges within the set $\{2000, 3000, 4000, 5000, 6000\}$. Supplementary Table S1 details the number of cells and genes for each synthetic dataset numbered from 1 to 30.

For those real-world datasets, they have six real single-cell RNA-seq datasets including Buettner, Deng, Ginhoux, Pollen, Ting, and Treutlin (detailed in Supplementary Section II). We summarize their characteristics of those datasets in Table I. From Table I, the number of cell is ranged from 80 to 251; the minimum number of genes is 959 while the maximum number of genes is 14805; and the number of cell types varies from 3 to 11.

TABLE I: Characteristics of six real single-cell RNA-seq datasets

Dataset	Cells	Genes	Cell types
Buettner	182	8989	3
Deng	135	12548	7
Ginhoux	251	11834	3
Pollen	249	14805	11
Ting	114	14405	5
Treutlin	80	959	5

B. Parameter Settings and Evaluation Metrics

In EMDC, the population size (N) is set to 200, the crossover rate (*CR*), and the mutation probability (*MP*) are set to 0.8 and 0.1, respectively (discussed in Section 3.10). To guarantee the fairness of the comparison, 1000 objective function evaluations (*FEs*) are taken as the stopping criterion [42] for each single-cell RNA-seq dataset. The variables used in this study are summarized in Supplementary Tables S2 and S3. In addition, to exclude contingency factor, we provide the average *NMI* and *ARI* under 30 independent runs on each dataset. *NMI* and *ARI* are adopted as evaluation metrics, which are detailed in Supplementary Section III.

C. Baseline Methods

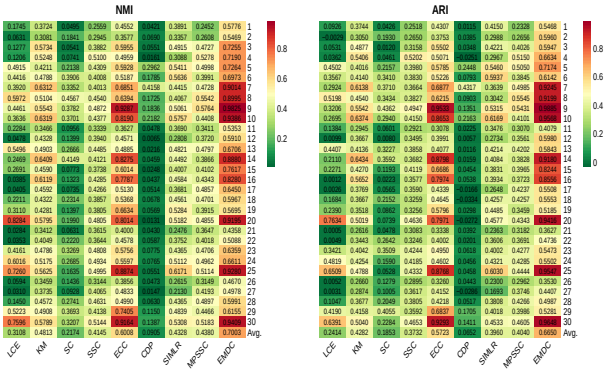
From the perspective of clustering, to rigorously assess the performance of EMDC, we compare it with eight clustering algorithms, including Linked-based Cluster Ensemble (LCE) [43], K-means clustering algorithm (KM) [34], Spectral Clustering (SC) [35], Sparse Spectral Clustering (SSC) [35], Entropy-based Consensus Clustering (ECC) [44], Clustering by Fast Search and Find of Density Peaks (CDP) [36], Single-Cell Interpretation via Multikernel Learning (SIMLR) [8], and Spectral clustering based on learning similarity matrix (MPSSC) [45]. Different clustering algorithms indicate different algorithmic paradigms. LCE employs the ensemble technique for clustering, KM is a simple clustering method that has been frequently used in clustering data, SC and SSC are clustering algorithms based on the spectral graph theory, ECC is an ensemble clustering method that applies the entropy-based utility function to merge the basic clusterings into a consensus clustering, CDP adopts the density peaks to explore the cluster centers, SIMLR learns from the data to obtain the similarity measure between samples, and MPSSC applies the sparse structure on the target matrix for clustering. The source codes of baseline algorithms are given below. LCE is from https://www.jstatsoft.org/article/view/v036i09; KM is from Matlab Library; SC and SSC are from https://github.com/ArrowLuo/Spectral_cluster_matlab and https://github.com/ishspsy/project/tree/master/MPSSC/SparseSC respectively; ECC is from http://scholar.harvard.edu/yyl/ecc; CDP is from http://people.sissa.it/~laio/Research/Clustering_source_code/cluster_dp.tgz; SIMLR can be found from https:

//github.com/ishspsy/project/tree/master/MPSSC/SIMLR; and MPSSC is from https://github.com/ishspsy/project/tree/master/MPSSC/Code.

From the perspective of multiobjective optimization, three multiobjective algorithms, including the Non-dominated Sorting Genetic Algorithm III (NSGAIII) [46], Multi-Objective Evolutionary Algorithm based on Decomposition (MOEA/D) [47], and Evolutionary Multiobjective Ensemble Pruning Algorithm (EMEP) [48] are exploited as the comparative algorithms. Each algorithm indicates an algorithmic paradigm. NSGAIII is a reference-point-based many-objective evolutionary algorithm, MOEA/D is a multiobjective technique based on decomposition, while EMEP is an evolutionary multiobjective ensemble pruning algorithm to enable the number of the chosen basic partition clusters minimized. Moreover, the time complexity of them is summarized in Supplementary Table S4.

To make statistical comparisons, we calculate the paired Wilcoxon test to demonstrate the difference between the proposed EMDC and the other comparative algorithm with a significant level 0.05. Two symbols H_1 and H_0 are used to indicate the difference between the paired algorithms. H_1 signifies that there is significant difference between them while H_0 represents that EMDC performs statistically comparable to the other one.

D. Evaluation on Synthetic Single-cell RNA-seq Data



(a) (b)

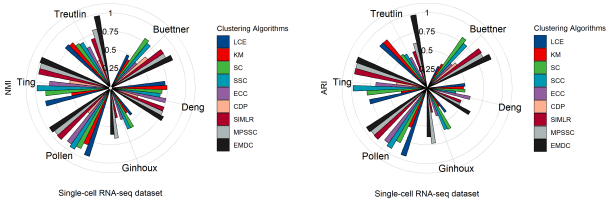
Fig. 3: Performance comparisons of different clustering algorithms measured by *NMI* (a) and *ARI* (b) on thirty synthetic single-cell RNA-seq datasets.

We conduct performance comparisons of the proposed EMDC and eight clustering algorithms including LCE, KM, SC, SSC, ECC, CDP, SIMLR, and MPSSC on thirty synthetic single-cell RNA-seq datasets. The experimental results measured by *NMI* and *ARI* are tabulated in Supplementary Tables S5 and S6. In these two tables, the best result on each dataset are highlighted in bold, and the statistical results obtained by the paired Wilcoxon test and the average values can be found in the last two rows. Besides, the performance comparison of EMDC and other clustering algorithms on each synthetic single-cell RNA-seq dataset is illustrated in Fig. 3.

The results in Supplementary Table S5 reveal that EMDC can reach the best *NMI* values for almost all those synthetic single-cell RNA-seq datasets. Compared with ECC, EMDC achieves the best *NMI* values on two datasets, while other clustering algorithms are all inferior to EMDC on those thirty single-cell RNA-seq datasets. For the average values, EMDC increases the average *NMI* values over other comparative algorithms at most 61% compared with that achieved by CDP and at least 10% compared with that obtained by ECC. Meanwhile, as demonstrated in Supplementary Table S6, EMDC obtains the best *ARI* values in 28 out of 30 cases over all synthetic single-cell RNA-seq datasets. Moreover, EMDC can achieve 66.50% in *ARI* while the comparative method ECC can provide 57.23%. The statistical results in those tables show that there are significant differences between EMDC and other clustering algorithms (p-value < 0.05). Besides, from Fig. 3, we can also observe that EMDC shows better performance than other clustering algorithms across almost all the datasets with respect of both *NMI* and *ARI*. In summary, it can conclude that the proposed EMDC demonstrates its robust performance in identifying cell types from single-cell RNA-seq data.

E. Evaluation on Real Single-cell RNA-seq Data

In this section, we compare the performance of EMDC and the other eight algorithms, LCE, KM, SC, SSC, ECC, CDP, SIMLR, and MPSSC on six real single-cell RNA-seq datasets to further demonstrate EMDC's effectiveness. We use *NMI* and *ARI* to measure the consistency between the predicted label and the ground truth label. The experimental results evaluated by *NMI* and *ARI* are illustrated in Supplementary Tables S7 and S8 respectively. Particularly, we provide the average values and the paired Wilcoxon test results in the last two rows of those tables. In addition, the superior clustering performance of EMDC is also expected from Fig. 4. The objective space visualization of EMDC on those real single-cell RNA-seq datasets is summarized in Supplementary Fig. S1.



(a) (b)

Fig. 4: Performance comparisons of different clustering algorithms measured by *NMI* (a) and *ARI* (b) on six single-cell RNA-seq datasets.

In terms of performance comparisons measured by *NMI*, as shown in Table II, EMDC consistently outperforms the existing methods on five datasets, including Buettner, Deng, Pollen, Ting, and Treutlin. The overall improvements fulfilled by EMDC in the average *NMI* across the six single-cell RNA-seq datasets exceed 4.81% over other clustering methods. In particular, EMDC increases the average *NMI* with a remarkably large percentage 48% over CDP. The *NMI* value on the Treutlin dataset shows remarkably better performance than other clustering algorithms. For the Ginhoux dataset, MPSSC behaves better than EMDC. The statistical results show that the differences between EMDC and other clustering algorithms are significantly different (p-value < 0.05) except for MPSSC.

In terms of performance comparisons measured by *ARI*, Table III demonstrates that EMDC performs better than seven

other algorithms, including LCE, KM, SC, SSC, ECC, CDP, and SIMLR. For MPSSC, EMDC has higher *ARI* values than it on five single-cell RNA-seq datasets except the Ginhoux dataset. In particular, for the Treutlin dataset, the largest *ARI* increase is 82% when EMDC is compared with CDP while the smallest increase is 14.55% between EMDC and LCE. The overall *ARI* improvement over other clustering methods is at least 9.7% fulfilled by EMDC over MPSSC. Meanwhile, as observed from the statistical results, EMDC is significantly different from LCE, KM, SC, SSC, ECC, CDP, and SIMLR.

Among those six single-cell RNA-seq datasets, two datasets Buettner and Treutlin are analyzed. The Buettner dataset contains 182 embryonic stem cells and 8989 genes with three cell types (G1, S, and G2M) based on the sorting of Hoechst stained cell area of flow cytometry distribution. Fig. 5 (a) visualizes the heatmap of that dataset using the similarity distance and the obtained labels by EMDC. The Treutlin dataset consists of 80 single distal lung epithelial cells and 959 genes with five cell types (AT1, AT2, ciliated, Clara, and BP) based on the existence of canonical marker genes, assigning to alveolar and bronchiolar lineages. Fig. 5 (b) shows the heatmap visualization of the Treutlin dataset. As depicted in Fig. 5 (a-b), we can find that EMDC can identify the true data structure by significant margins. In addition, *t*-distributed stochastic neighbour embedding algorithm (t-SNE) [16] is employed to project the data into two dimensions for visualization, which can intuitively demonstrate the hidden structures in the data. The visualization is implemented by adding the single-cell RNA-seq data with the obtained labels by EMDC and the ground truth labels. Fig. 5 (c-f) depicts the 2-D space visualization for those two datasets.

Based on the analysis above, we conclude that EMDC generally outperforms other clustering methods in uncovering cell-to-cell similarity and dissimilarity structures.

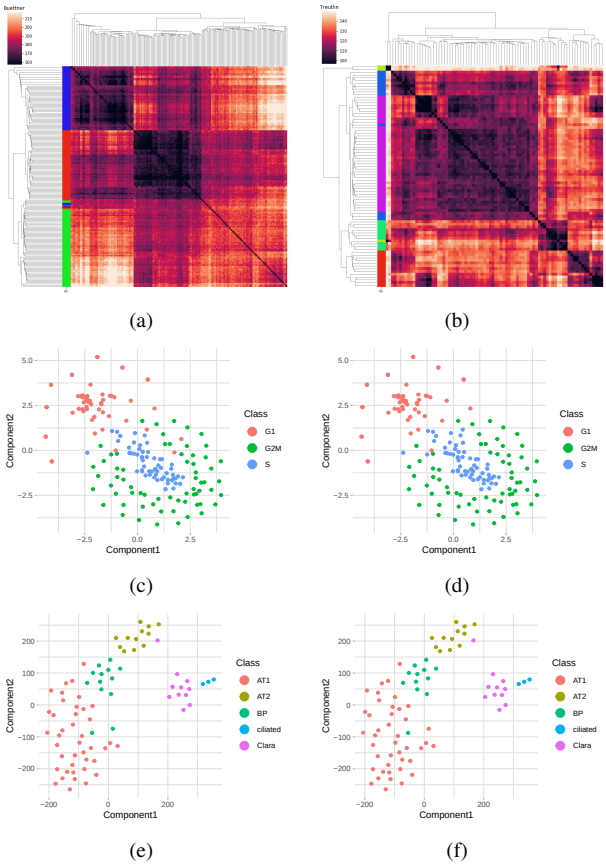


Fig. 5: (a) Heatmap visualizes the Buettner dataset with three cell types. (b) Heatmap visualizes the Treutlin dataset with five cell types. (c) 2-D visualization for the Buettner dataset with the truth label. (d) 2-D visualization for the Buettner dataset with the obtained label by EMDC. (e) 2-D visualization for the Treutlin dataset with the truth label. (f) 2-D visualization for the Treutlin dataset with the obtained label by EMDC. Different colors mark different cell types.

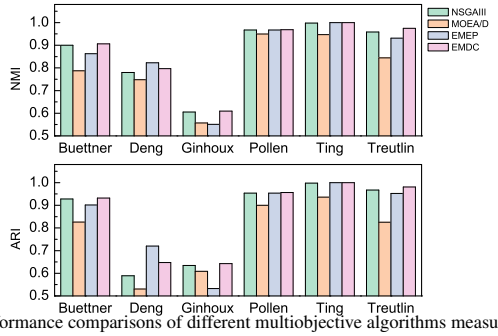


Fig. 6: Performance comparisons of different multiobjective algorithms measured by *NMI* and *ARI* on six single-cell RNA-seq datasets. The horizontal axis is the single-cell RNA-seq dataset while the vertical axis is the average *NMI* and *ARI* obtained by different multiobjective algorithms.

F. Multiobjective Optimization Methodology Comparisons

In this section, EMDC is compared with three multiobjective optimization algorithms including NSGAIII, MOEA/D, and EMEP on six real single-cell RNA-seq datasets to demonstrate its performance. Each algorithm conducts 30 runs on each dataset. The comparative results measured by *NMI* and *ARI* are tabulated in Supplementary Tables S9 and S10, respectively. Moreover, Fig. 6 depicts the performance of each algorithm. As illustrated in Supplementary Tables S9 and S10, and Fig. 6, for the Deng dataset, EMEP achieves the best *NMI* and *ARI* results; for other datasets, EMDC performs better than other multiobjective optimization algorithms. In addition, EMDC increases the average *NMI* value across all the single-cell RNA-seq datasets by 0.78%, 7.04%, and 2% over NSGAIII, MOEA/D, and EMEP, respectively. Similarly, EMDC improves the average *ARI* over NSGAIII, MOEA/D, and EMEP by 1.48%, 8.87%, and 1.65%. To sum up, EMDC based on hypervolume reflects its advantages in clustering the single-cell RNA-seq datasets.

G. Comparative Analysis of Dimension Reduction Methods

In this study, we analyze the effectiveness of deep autoencoder in our proposed EMDC. Four widely-used unsupervised dimension reduction methods, including Non-negative Matrix Factorization (NMF) [49], Independent Component Analysis (ICA) [50], Principal Component Analysis (PCA) [14], and Locality Preserving Projections (LPP) [51] are compared with the deep autoencoder used in EMDC. Among those dimension reduction methods, NMF is a matrix decomposition method to realize the reduction of non-linear dimensions; ICA is a linear dimension reduction method to search latent factors or components that satisfy statistical independence and non-Gaussian from multidimensional statistical data; PCA is also a simple linear dimension reduction method that maintains the characteristics of the largest variance contribution in the dataset; and LPP is a method to reduce the dimension while

retaining the local neighborhood structure of the sample in the space. In the experiment, each dimension reduction method is employed to project the single-cell RNA-seq data onto different low dimensions from 20 to 200 sequentially. We name each method as EMDC$_{NMF}$, EMDC$_{ICA}$, EMDC$_{PCA}$, and EMDC$_{LPP}$ respectively. The performance of each method is measured by *NMI* and *ARI* for 30 independent runs. The experimental results are summarized in Supplementary Tables S11 and S12. Moreover, Fig. 7 shows the *NMI* results of those comparative methods to clearly demonstrate the superiority of the deep autoencoder in EMDC. As observed in Fig. 7, EMDC$_{PCA}$ can achieve slightly higher *NMI* values than EMDC on the Deng and Pollen datasets. Meanwhile, EMDC$_{PCA}$ performs better than EMDC regarding *ARI* on the Pollen dataset. For EMDC$_{NMF}$, EMDC$_{ICA}$, and EMDC$_{LPP}$, compared with EMDC, they cannot perform the best results on all the datasets. Moreover, EMDC with the deep autoencoder significantly outperforms other methods in terms of the average results across those six single-cell RNA-seq datasets. In summary, we observed that the deep autoencoder can enhance the EMDC's performance of cell type identification.

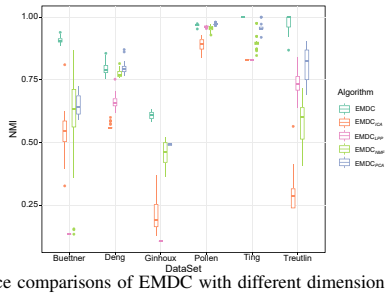


Fig. 7: Performance comparisons of EMDC with different dimension reduction methods measured by *NMI* on six single-cell RNA-seq datasets. The horizontal axis is the single-cell RNA-seq dataset while the vertical axis is the average *NMI* obtained by EMDC with different dimension reduction methods.

H. Different Objective Function Subsets

TABLE II: Performance comparisons of EMDC with different combinations of objective functions on six single-cell RNA-seq datasets measured by *NMI* and *ARI*.

Objective Functions	Db+Nc		Dunn+Nc		Db+Dunn+Nc	
DataSet	*NMI*	*ARI*	*NMI*	*ARI*	*NMI*	*ARI*
Buettner	0.8777	0.9068	0.8511	0.8867	**0.9060**	**0.9318**
Deng	0.7911	0.6380	0.7838	0.6157	**0.7966**	**0.6474**
Ginhoux	0.5936	0.6379	0.6065	0.6350	**0.6094**	**0.6428**
Pollen	0.9697	0.9581	0.9618	0.9437	**0.9684**	**0.9563**
Ting	0.9938	0.9933	0.9915	0.9926	**1.0000**	**1.0000**
Treutlin	0.9584	0.9711	0.9181	0.9182	**0.9746**	**0.9808**
Avg.	0.8641	0.8509	0.8521	0.8320	**0.8758**	**0.8599**

This section is designed to demonstrate the effectiveness of the proposed objective function combination under the multiobjective framework in EMDC. In our experiment, we compare EMDC with two different combinations of objective functions on six real single-cell RNA-seq datasets. It is noted that the objective function Nc is embraced in all the compared objective function sets because of the pruning characteristic of EMDC. The experimental results measured by *NMI* and *ARI* are tabulated in Table IV. As observed in Table IV, our proposed algorithm is superior to others on five single-cell RNA-seq datasets including Buettner, Deng, Ginhoux, Ting, and Treutlin. For the Pollen dataset, EMDC under the objective function set containing Db and Nc performs slightly better than EMDC. In general, the experimental results demonstrate that the proposed objective function subsets can better capture the characteristics of the single-cell RNA-seq data and facilitate to enhance the optimization ability of EMDC.

I. Parameter Analysis

1) Sensitivity of Crossover Rate (CR): We discuss the sensitivity of *CR* in this section. In the experiment, *CR* ranges within the set $\{0.2, 0.4, 0.6, 0.8, 1\}$. The experimental results are measured by the average *NMI* for 30 independent runs on each single-cell RNA-seq dataset. Performance of EMDC with different *CR* settings are summarized in Supplementary Fig. S2. Supplementary Fig. S2 reveals that EMDC is insensitive to the crossover rate, demonstrating the robustness of EMDC.

2) Effect of Mutation Probability (MP): In this section, the robustness of the proposed algorithm is evaluated with varying mutation probabilities chosen from the set $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. EMDC with each parameter setting runs 30 times on each single-cell RNA-seq dataset. The performance is measured by *NMI* and summarized in Supplementary Fig. S3. As depicted in Supplementary Fig. S3, EMDC can obtain consistent performance under different mutation probabilities. Since *MP*=0.1 provides slightly better results than other *MP* settings, we set *MP*=0.1 in EMDC.

3) Convergence Behavior: In this section, a convergence analysis is conducted to demonstrate EMDC's convergence behaviour. We vary the number of objective function evaluations (*FEs*) from 50 to 3000 for EMDC on those real single-cell RNA-seq datasets. The performance is evaluated by the average *NMI* and *ARI* for 30 independent runs across those datasets. The convergence trajectory of EMDC is summarized in Fig. 8. As depicted in Fig. 8, the overall performance of EMDC increases as the number of *FEs* increases. In addition, the improvement of the performance tends to reach a steady state for EMDC after 1000 *FEs*. It implies that *FEs=1000* is a reasonable setting in EMDC for the good convergence.

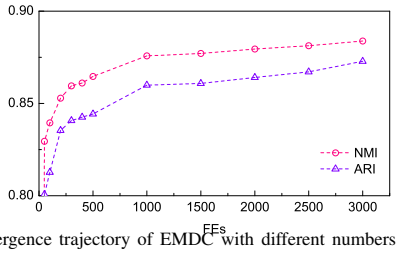


Fig. 8: The convergence trajectory of EMDC with different numbers of *FEs* measured by the average *NMI* and *ARI* across six single-cell RNA-seq datasets. The horizontal axis is the number of *FEs* and the vertical axis is the average *NMI* and *ARI*.

J. Extended Analysis and Comparisons

TABLE III: Performance comparisons of EMDC with different ensemble construction methods on six single-cell RNA-seq datasets measured by *NMI* and *ARI*

Algorithm	EMDC-KM++		EMDC-AL		EMDC-CL		EMDC-SL		EMDC	
DataSet	*NMI*	*ARI*	*NMI*	*ARI*	*NMI*	*ARI*	*NMI*	*ARI*	*NMI*	*ARI*
Buettner	0.8423	0.8767	0.6137	0.4599	0.8208	0.8485	0.1793	0.0586	**0.9060**	**0.9318**
Deng	**0.8132**	0.6469	0.7891	**0.6652**	0.7779	0.5757	0.5594	0.3112	0.7966	0.6474
Ginhoux	0.5947	0.6362	0.5011	0.5054	0.5505	0.6007	0.1151	0.0098	**0.6094**	**0.6428**
Pollen	**0.9686**	**0.9565**	0.8787	0.6968	0.9062	0.7730	0.5490	0.2204	0.9684	0.9563
Ting	0.9919	0.9932	0.9291	0.9138	**1.0000**	**1.0000**	0.8311	0.7658	**1.0000**	**1.0000**
Treutlin	0.8862	0.9033	0.8365	0.6609	0.9636	0.9647	0.7005	0.5315	**0.9746**	**0.9808**
Avg.	0.8495	0.8355	0.7580	0.6503	0.8365	0.7938	0.4891	0.3162	**0.8758**	**0.8599**

1) Ensemble Construction Method: Basic clustering members play crucial roles in promoting the effectiveness of the ensemble algorithm, since an ensemble will take advantage of each member to enhance the generalisability of the algorithm. This section is devoted to discuss whether the

TABLE IV: Performance comparisons of EMDC with different base clustering algorithms on six single-cell RNA-seq datasets measured by *NMI* and *ARI*

Algorithm	$EMDC_1$		$EMDC_2$		$EMDC_3$		EMDC	
DataSet	*NMI*	*ARI*	*NMI*	*ARI*	*NMI*	*ARI*	*NMI*	*ARI*
Buettner	0.8975	0.9245	0.9057	**0.9318**	0.1361	0.0157	**0.9060**	**0.9318**
Deng	**0.7990**	**0.6533**	0.7948	0.6021	0.5594	0.3112	0.7966	0.6474
Ginhoux	0.5996	0.6339	**0.6128**	0.6398	0.1076	-0.0019	0.6094	**0.6428**
Pollen	0.9695	0.9583	**0.9703**	**0.9608**	0.5437	0.2056	0.9684	0.9563
Ting	0.9992	0.9993	0.9865	0.9848	0.8311	0.7658	**1.0000**	**1.0000**
Treutlin	0.9561	0.9645	0.9625	0.9678	0.2405	0.0442	**0.9746**	**0.9808**
Avg.	0.8702	0.8556	0.8721	0.8479	0.4031	0.2234	**0.8758**	**0.8599**

ensemble construction method in EMDC (K-means clustering algorithm) outperforms other ensemble construction methods. Four simple clustering algorithms including K-means++ [52], hierarchical clustering with average-linkage (AL), complete-linkage (CL), and single-linkage (SL) [53], are compared against K-means clustering in EMDC to generate some basic clustering members. Each method is named EMDC-KM++, EMDC-AL, EMDC-CL, and EMDC-SL accordingly. In the experiment, each method runs over 30 times for the fairness of comparison, while *NMI* and *ARI* are considered to evaluate the performance of each method. The experimental results are summarized in Table V. From Table V, EMDC achieves the best performance in terms of *NMI* on the Buettner, Ginhoux, Ting, and Treutlin datasets whereas EMDC-KM++ can provide the best performance on the Deng and Pollen datasets. Similarly, with respect to *ARI* in Table V, although EMDC-KM++ obtains slightly better *ARI* result on the Pollen dataset, EMDC outperforms EMDC-KM++ on all of the other datasets. For EMDC-AL and EMDC-CL, they provide the best *ARI* results on the Deng and Ting datasets. For EMDC-SL, it cannot obtain the best *NMI* and *ARI* results on all the real single-cell RNA-seq datasets. In conclusion, our proposed EMDC exhibits overall the best performance, which reflects that the importance of choosing the suitable ensemble construction method for clustering those single-cell RNA-seq datasets.

2) Base Clustering Algorithm Discussion: To demonstrate the effectiveness of those base clustering algorithms for the proposed EMDC, we discuss the impact of each base clustering algorithm (K-means, SC, and CDP) on EMDC. The experimental results are reported in Table VI. We represent EMDC with K-means, SC, and CDP as $EMDC_1$, $EMDC_2$, and $EMDC_3$, respectively. As observed from the tables, with respect to *NMI*, for the Buettner, Ting, and Treutlin datasets, EMDC obtains the best *NMI* results. For the Ginhoux and Pollen datasets, $EMDC_2$ can provide the best performance. For the Deng dataset, $EMDC_1$ achieves the highest *NMI* value. For *ARI*, EMDC provides the best performance on four out of those six datasets. $EMDC_2$ and EMDC obtain a similar performance for the Buettner dataset. $EMDC_1$ generates slightly better *ARI* result than EMDC on the Deng dataset. In addition, the average *NMI* and *ARI* across those six datasets obtained by EMDC are higher than those obtained by EMDC under a single base clustering algorithm. In particular, EMDC improves the average *NMI* and *ARI* by 47% and 63% over $EMDC_3$. The above analysis suggests that EMDC with a combination of those three base clustering algorithms performs better than that under a single base clustering algorithm. It is appropriate for EMDC to select the clustering algorithm evolutionarily from those base clustering algorithms instead of employing a single base clustering algorithm in grouping various cell types from different single-cell RNA-seq data.

IV. Conclusion

In this study, we propose an evolutionary multiobjective deep clustering (EMDC) algorithm that integrates data preprocessing, data dimensionality reduction, and data clustering to identify cell types from the single-cell RNA-seq data. First, the differential gene expression analysis is utilized to select a number of genes from the high-dimensional data in the original space. Then, deep autoencoder is employed to capture the inherent structure of those preprocessed data efficiently, learning different latent feature representations with multiple low-dimensional spaces. After that, the basic clustering algorithm is applied across all those latent feature representations to generate an ensemble including a set of basic clusterings. In EMDC, we encode the population to generate several ensembles with different basic clusterings. Then, the population is evolved under the multiobjective optimization. Three objective functions, including Db, Dunn, and the number of base clusterings are designed to guide the evolution, capturing various characteristics of the evolving clustering in EMDC. To validate the effectiveness of the proposed algorithm, we apply thirty synthetic single-cell RNA-seq datasets and six real single-cell RNA-seq datasets for our proposed EMDC algorithm and other comparative approaches that contain several state-of-the-art clustering methods and multiobjective optimization algorithms. Besides, extensive experiments are performed to demonstrate the robustness of EMDC.

In the future, we would like to design other consensus functions for the single-cell RNA-seq data under the multiobjective optimization framework. Moreover, we are interested in applying the proposed EMDC to other high-dimensional molecular data.

References

[1] C. Ziegenhain, B. Vieth, S. Parekh, B. Reinius, A. Guillaumet-Adkins, M. Smets, H. Leonhardt, H. Heyn, I. Hellmann, and W. Enard, "Comparative analysis of single-cell rna sequencing methods," *Molecular cell*, vol. 65, no. 4, pp. 631–643, 2017.

[2] A. Peyvandipour, A. Shafi, N. Saberian, and S. Draghici, "Identification of cell types from single cell data using stable clustering," *Scientific reports*, vol. 10, no. 1, pp. 1–12, 2020.

[3] V. Y. Kiselev, T. S. Andrews, and M. Hemberg, "Challenges in unsupervised clustering of single-cell rna-seq data," *Nature Reviews Genetics*, vol. 20, no. 5, pp. 273–282, 2019.

[4] Y. Gan, N. Li, G. Zou, Y. Xin, and J. Guan, "Identification of cancer subtypes from single-cell rna-seq data using a consensus clustering method," *BMC medical genomics*, vol. 11, no. 6, pp. 65–72, 2018.

[5] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[6] P. Lin, M. Troup, and J. W. Ho, "Cidr: Ultrafast and accurate clustering through imputation for single-cell rna-seq data," *Genome biology*, vol. 18, no. 1, p. 59, 2017.

[7] J. H. Levine, E. F. Simonds, S. C. Bendall, K. L. Davis, D. A. El-ad, M. D. Tadmor, O. Litvin, H. G. Fienberg, A. Jager, E. R. Zunder *et al.*, "Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis," *Cell*, vol. 162, no. 1, pp. 184–197, 2015.

[8] B. Wang, J. Zhu, E. Pierson, D. Ramazzotti, and S. Batzoglou, "Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning," *Nature methods*, vol. 14, no. 4, p. 414, 2017.

[9] X. Li, S. Zhang, and K.-C. Wong, "Single-cell rna-seq interpretations using evolutionary multiobjective ensemble pruning," *Bioinformatics*, vol. 35, no. 16, pp. 2809–2817, 2019.

[10] S. Asur, D. Ucar, and S. Parthasarathy, "An ensemble framework for clustering protein–protein interaction networks," *Bioinformatics*, vol. 23, no. 13, pp. i29–i40, 2007.

[11] D. Huang, C.-D. Wang, and J.-H. Lai, "Locally weighted ensemble clustering," *IEEE transactions on cybernetics*, vol. 48, no. 5, pp. 1460–1473, 2017.

[12] Y. Yang, R. Huh, H. W. Culpepper, Y. Lin, M. I. Love, and Y. Li, "Safe-clustering: Single-cell aggregated (from ensemble) clustering for single-cell rna-seq data," *Bioinformatics*, vol. 35, no. 8, pp. 1269–1277, 2019.

[13] J. Shin, D. A. Berg, Y. Zhu, J. Y. Shin, J. Song, M. A. Bonaguidi, G. Enikolopov, D. W. Nauen, K. M. Christian, G.-l. Ming *et al.*, "Single-cell rna-seq with waterfall reveals molecular cascades underlying adult neurogenesis," *Cell stem cell*, vol. 17, no. 3, pp. 360–372, 2015.

[14] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[15] D. Grün, A. Lyubimova, L. Kester, K. Wiebrands, O. Basak, N. Sasaki, H. Clevers, and A. Van Oudenaarden, "Single-cell messenger rna sequencing reveals rare intestinal cell types," *Nature*, vol. 525, no. 7568, pp. 251–255, 2015.

[16] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[17] Y.-X. Wang and Y.-J. Zhang, "Nonnegative matrix factorization: A comprehensive review," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 6, pp. 1336–1353, 2012.

[18] D. Wang and J. Gu, "Vasc: dimension reduction and visualization of single cell rna sequencing data by deep variational autoencoder," *bioRxiv*, p. 199315, 2017.

[19] A. Tangherloni, F. Ricciuti, D. Besozzi, P. Liò, and A. Cvejic, "scaespy: a unifying tool based on autoencoders for the analysis of single-cell rna sequencing data," *bioRxiv*, p. 727867, 2019.

[20] L. Chen, W. Wang, Y. Zhai, and M. Deng, "Single-cell transcriptome data clustering via multinomial modeling and adaptive fuzzy k-means algorithm," *Frontiers in Genetics*, vol. 11, p. 295, 2020.

[21] K. Smyth Gordon, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, pp. 1–25, 2004.

[22] R. Silipo, I. Adae, A. Hart, and M. Berthold, "Seven techniques for dimensionality reduction," *White Paper by KNIME. com AG*, pp. 1–21, 2014.

[23] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[26] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 12, pp. 1650–1654, 2002.

[27] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, "Validity index for crisp and fuzzy clusters," *Pattern recognition*, vol. 37, no. 3, pp. 487–501, 2004.

[28] K. Deb, R. B. Agrawal *et al.*, "Simulated binary crossover for continuous search space," *Complex systems*, vol. 9, no. 2, pp. 115–148, 1995.

[29] K. Deb and S. Tiwari, "Omni-optimizer: A generic evolutionary algorithm for single and multi-objective optimization," *European Journal of Operational Research*, vol. 185, no. 3, pp. 1062–1087, 2008.

[30] J. Bader and E. Zitzler, "Hype: An algorithm for fast hypervolume-based many-objective optimization," *Evolutionary Computation*, vol. 19, no. 1, pp. 45–76, 2011.

[31] S. Klink, P. Reuther, A. Weber, B. Walter, and M. Ley, "Analysing social networks within bibliographical data," in *International Conference on Database and Expert Systems Applications*. Springer, 2006, pp. 234–243.

[32] P. Calado, M. Cristo, M. A. Gonçalves, E. S. de Moura, B. Ribeiro-Neto, and N. Ziviani, "Link-based similarity measures for the classification of web documents," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 2, pp. 208–221, 2006.

[33] N. Iam-On, T. Boongeon, S. Garrett, and C. Price, "A link-based cluster ensemble approach for categorical data clustering," *IEEE Transactions on knowledge and data engineering*, vol. 24, no. 3, pp. 413–425, 2010.

[34] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl *et al.*, "Constrained k-means clustering with background knowledge," in *Icml*, vol. 1, 2001, pp. 577–584.

[35] U. V. Luxburg, "A tutorial on spectral clustering," *Statistics & Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[36] A. Rodriguez and A. Laio, "Machine learning. clustering by fast search and find of density peaks." *Science*, vol. 344, no. 6191, p. 1492, 2014.

[37] K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan, "A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: Nsga-ii," in *International*

conference on parallel problem solving from nature. Springer, 2000, pp. 849–858.
[38] C. Igel, N. Hansen, and S. Roth, "Covariance matrix adaptation for multi-objective optimization," *Evolutionary computation*, vol. 15, no. 1, pp. 1–28, 2007.
[39] A. Mukhopadhyay, U. Maulik, and S. Bandyopadhyay, "A survey of multiobjective evolutionary clustering," *ACM Computing Surveys (CSUR)*, vol. 47, no. 4, pp. 1–46, 2015.
[40] X. Li and K.-C. Wong, "Evolutionary multiobjective clustering and its applications to patient stratification," *IEEE transactions on cybernetics*, vol. 49, no. 5, pp. 1680–1693, 2018.
[41] A. T. Assefa, J. Vandesompele, and O. Thas, "Spsimseq: semi-parametric simulation of bulk and single-cell rna-sequencing data," *Bioinformatics*, vol. 36, no. 10, pp. 3276–3278, 2020.
[42] X. Li and K.-C. Wong, "Multiobjective patient stratification using evolutionary multiobjective optimization," *IEEE journal of biomedical and health informatics*, vol. 22, no. 5, pp. 1619–1629, 2017.
[43] N. Iam-On, T. Boongoen, and S. Garrett, "Lce: a link-based cluster ensemble method for improved gene expression data analysis," *Bioinformatics*, vol. 26, no. 12, pp. 1513–1519, 2010.
[44] H. Liu, R. Zhao, H. Fang, F. Cheng, Y. Fu, and Y. Y. Liu, "Entropy-based consensus clustering for patient stratification." *Bioinformatics*, vol. 33, no. 17, 2017.
[45] S. Park and H. Zhao, "Spectral clustering based on learning similarity matrix," *Bioinformatics*, vol. 34, no. 12, pp. 2069–2076, 2018.
[46] K. Deb and H. Jain, "An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part i: solving problems with box constraints," *IEEE transactions on evolutionary computation*, vol. 18, no. 4, pp. 577–601, 2013.
[47] Q. Zhang and L. Hui, "Moea/d: A multiobjective evolutionary algorithm based on decomposition," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 6, pp. 712–731, 2007.
[48] X. Li, S. Zhang, and K.-C. Wong, "Single-cell rna-seq interpretations using evolutionary multiobjective ensemble pruning," *Bioinformatics*, 2018.
[49] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.
[50] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.
[51] X. He and P. Niyogi, "Locality preserving projections," in *Advances in neural information processing systems*, 2004, pp. 153–160.
[52] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," Stanford, Tech. Rep., 2006.
[53] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: A survey," *IEEE Transactions on Knowledge and Data Engineering*, no. 11, pp. 1370–1386, 2004.

Yunhe Wang received the BEng, MPhil, and PhD degrees in computer science from Northeast Normal University, Changchun, China in 2014, 2017, 2021, respectively. She is now a lecturer in the School of Artificial Intelligence, Hebei University of Technology. She also was a visiting PhD student at the Department of Computer Science and Informatics, De Montfort University, Leicester, UK. Her current research interests include evolutionary computation, swarm intelligence, and machine learning.

Chuang Bian is a postgraduate with School of Artificial Intelligence, Jilin University, Changchun, China. His current research interest is bioinformatics.

Ka-Chun Wong received the BEng degree, the MPhil degree in computer engineering from United College, Chinese University of Hong Kong in 2008, 2010. He received the PhD degree from the School of Computer Science, University of Toronto in 2014. He assumed his duty as an assistant professor at City University of Hong Kong in 2015. His research interests include bioinformatics, evolutionary computing, data mining, machine learning, and interdisciplinary research.

He was named the associate editor of BioData Mining in 2016. Besides, he has been the editorial board of Applied Soft Computing since 2016. It is worth noting that he has solely edited two books published by Springer and CRC Press, attracting thirty peer-reviewed books chapters around the world.

Xiangtao Li (M'15) received the BEng, MPhil, and PhD degrees in computer science from Northeast Normal University, Changchun, China in 2009, 2012, 2015, respectively. He is now a Professor in the School of Artificial Intelligence, Jilin University. He has more than fifty publications. His research interests include evolutionary computation, constrained optimization, bioinformatics, and computational biology.

Shengxiang Yang (Senior Member, IEEE) received the Ph.D. degree from Northeastern University, Shenyang, China, in 1999. He is currently a Professor of Computational Intelligence and the Director of the Centre for Computational Intelligence, School of Computer Science and Informatics, De Montfort University, Leicester, U.K. He has over 340 publications with an H-index of 60 according to Google Scholar. His current research interests include evolutionary computation, swarm intelligence, artificial neural networks, data mining and data stream mining, and relevant real-world applications. Prof. Yang serves as an Associate Editor/Editorial Board Member of a number of international journals, such as the *IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION*, *IEEE TRANSACTIONS ON CYBERNETICS*, *Information Sciences*, *Enterprise Information Systems*, and *CAAI Transactions on Intelligence Technology*.