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

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

This is a supplementary material of the paper "Multiobjective Deep Clustering and Its Applications in Single-cell RNA-seq Data", submitted to IEEE Transactions on Systems, Man and Cybernetics: Systems.

I. Time Complexity Analysis

In this section, the time complexity of the proposed algorithm EMDC is analyzed. For EMDC, it contains two components: the deep autoencoder and the evolutionary multiobjective optimization. The deep autoencoder costs $O(nI^2 + nmK)$, where n is the number of samples in the single-cell RNA-seq data, I is the number of neurons in hidden layers, m is the dimension of latent feature representation, and K is the number of clusters in the clustering. Due to $K \leq m \leq I$, the time complexity is approximated as $O(nI^2)$ [1]. For the evolutionary multiobjective optimization, the crossover and mutation operator cost $O(2N)$ in each iteration, where N is the population size. Then the consensus function and the base clustering algorithm are selected adaptively to calculate the objective functions of each individual. For the consensus function, it costs $O(n^2d)$ for the similarity computation, where d is the number of basic clusterings in the ensemble. For the base clustering algorithm, KM costs $O(tKdn)$, where t is a fixed number of iterations; CDP costs $O(n^2)$ by finding the cluster center; SC costs $O(n^3)$ for calculating the eigenvalues of the similarity matrix of the data. Therefore, the worst time complexity for N individuals is $O(N \times (n^2d + n^3))$. After that, the Pareto optimal approach that mainly lies in the update process costs $O(MN^2)$ [2], where M is the number of objective functions. Given T is the number of iterations for the optimization process, $2N \ll N(n^2d+n^3)$ holds, the time complexity of the evolutionary multiobjective optimization is $O(T(N(n^2d+n^3)+MN^2))$. In conclusion, the overall time complexity of EMDC is $O(nI^2 + TN(n^2d + n^3 + MN))$.

II. Data Collection

The Buettner dataset [3] is comprised of 182 embryonic stem cells and 8989 genes with three cell types (G1, S, and G2M) based on the Hoechst stained cell area sorting of flow cytometry distribution; the Deng dataset [4] has 135 single cells isolated from mouse embryos and 12548 genes with seven clusters (zygote, early 2-cell-stage, mid 2-cell-stage, late 2-cell-stage, 4-cell-stage, 8-cell-stage, and 16-cell-stage); the Ginhoux dataset [5] consists of 251 dendritic cell progenitors and 11834 genes under three cellular states that are Monocyte and Dendritic cell Progenitors, Common Dendritic cell Progenitors, and Pre-Dendritic Cells; the Pollen dataset [6] is comprised of 249 single cells corresponding to pluripotent cells, skin cells, blood cells, and neural cells, with 14805 genes, and divided into 11 cell populations; the Ting dataset [7] contains 114 pancreatic circulating tumor cells and 14405 genes and is divided into five clusters; the Treutlin dataset [8] contains 80 single distal lung epithelial cells and 959 genes with five clusters (AT2, AT1, ciliated, Clara, and BP) based on the presence of canonical marker genes.

III. Evaluation Metrics

To evaluate the predicted results, two most widely used evaluation metrics are adopted, which are Normalized Mutual Information (*NMI*) [9] and Adjusted Rand Index (*ARI*) [10]. *NMI* measures the similarity between two clustering results while *ARI* evaluates the consistency between two clustering results. Setting π_e as the obtained clustering and π_t as the ground truth clustering, they can be calculated as follows:

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)

$$NMI(\pi_e,\pi_t)=\frac{\sum_{i=1}^{n_e}\sum_{j=1}^{n_t} n_{ij}\log\frac{n_{ij}n}{n_e^i n_t^j}}{\sqrt{(\sum_{i=1}^{n_e} n_e^i\log\frac{n_e^i}{n})(\sum_{j=1}^{n_t} n_t^j\log\frac{n_t^j}{n})}} \tag{1}$$

$$ARI(\pi_e,\pi_t)=\frac{\sum_{i=1}^{n_e}\sum_{j=1}^{n_t}\binom{n_{ij}}{2}-\sum_{i=1}^{n_e}\binom{n_e^i}{2}\cdot\sum_{j=1}^{n_t}\binom{n_t^j}{2}/\binom{n}{2}}{\sum_{i=1}^{n_e}\binom{n_e^i}{2}/2+\sum_{j=1}^{n_t}\binom{n_t^j}{2}/2-\sum_{i=1}^{n_e}\binom{n_e^i}{2}\cdot\sum_{j=1}^{n_t}\binom{n_t^j}{2}/\binom{n}{2}} \tag{2}$$

where n_e, n_t are the numbers of clusters in π_e and π_t, respectively. n_e^i is the sample size in cluster i of π_e, n_t^j is the sample size in cluster j of π_t, and n_{ij} is the intersection sample size between cluster i of π_e and cluster j of π_t.

REFERENCES

[1] X. Guo, L. Gao, X. Liu, and J. Yin, "Improved deep embedded clustering with local structure preservation." in *IJCAI*, 2017, pp. 1753–1759.

[2] 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.

[3] B. Florian, K. N. Natarajan, C. F Paolo, P. Valentina, S. Antonio, F. J. Theis, S. A. Teichmann, J. C. Marioni, and S. Oliver, "Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells," *Nature Biotechnology*, vol. 33, no. 2, pp. 155–160, 2015.

[4] Q. Deng, D. Ramsköld, B. Reinius, and R. Sandberg, "Single-cell rna-seq reveals dynamic, random monoallelic gene expression in mammalian cells," *Science*, vol. 343, no. 6167, pp. 193–196, 2014.

[5] A. Schlitzer, V. Sivakamasundari, J. Chen, H. R. Sumatoh, J. Schreuder, J. Lum, B. Malleret, S. Zhang, A. Larbi, and F. Zolezzi, "Identification of cdc1- and cdc2-committed dc progenitors reveals early lineage priming at the common dc progenitor stage in the bone marrow," *Nature Immunology*, vol. 16, no. 7, pp. 718–728, 2015.

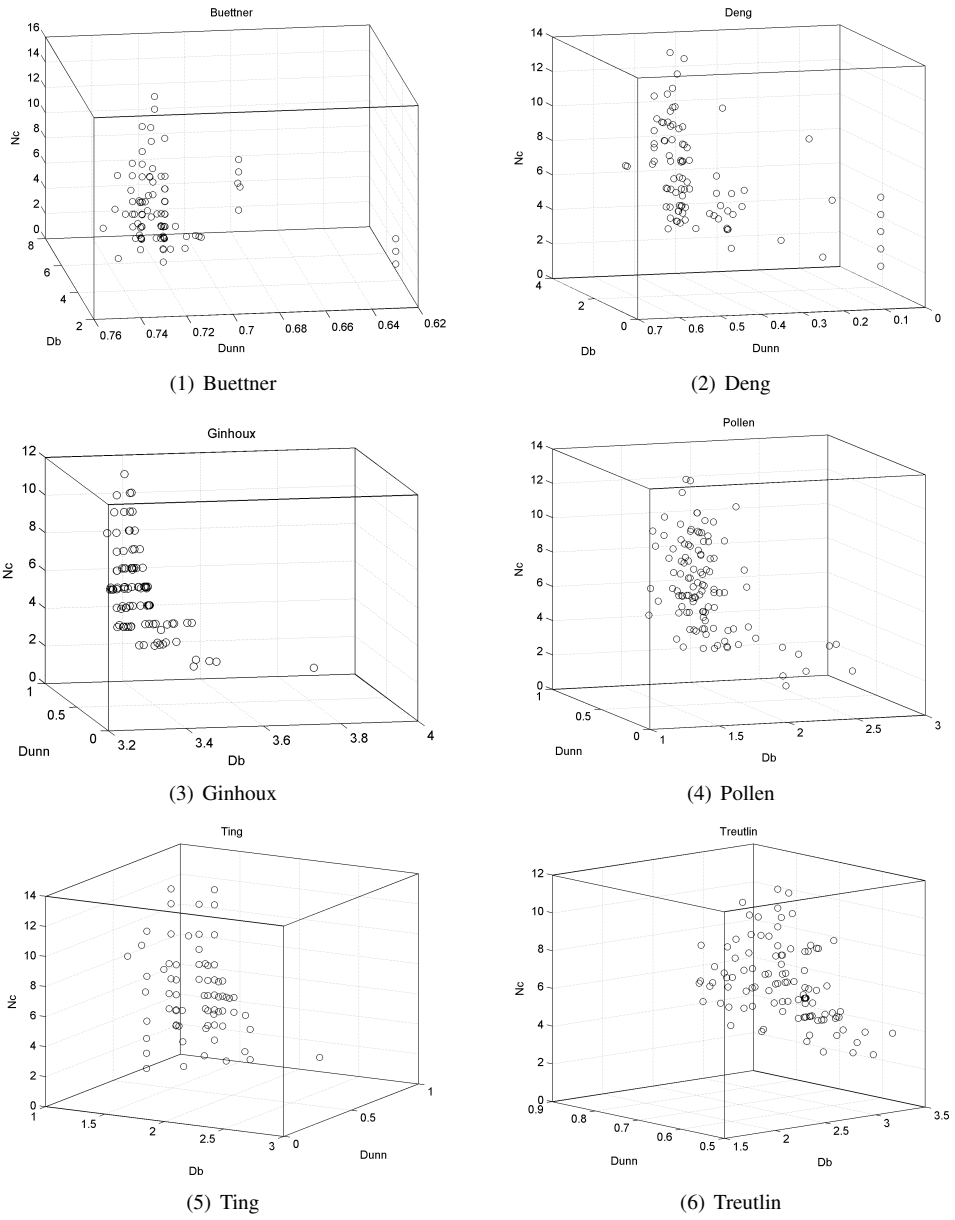
[6] A. A. Pollen, T. J. Nowakowski, S. Joe, W. Xiaohui, A. A. Leyrat, J. H. Lui, L. Nianzhen, S. Lukasz, F. Brian, and C. Peilin, "Low-coverage single-cell mrna sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex," *Nature Biotechnology*, vol. 32, no. 10, pp. 1053–1058, 2014.

[7] D. T. Ting, B. S. Wittner, L. Matteo, V. J. Nicole, A. M. Shah, D. T. Miyamoto, A. Nicola, B. Francesca, B. W. Brannigan, and X. Kristina, "Single-cell rna sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells," *Cell Reports*, vol. 8, no. 6, pp. 1905–1918, 2014.

[8] T. Barbara, D. G. Brownfield, A. R. Wu, N. F. Neff, G. L. Mantalas, E. F Hernan, T. J. Desai, M. A. Krasnow, and S. R. Quake, "Reconstructing lineage hierarchies of the distal lung epithelium using single-cell rna-seq," *Nature*, vol. 509, no. 7500, pp. 371–375, 2014.

[9] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," *Journal of machine learning research*, vol. 3, no. Dec, pp. 583–617, 2002.

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



(1) Buettner (2) Deng

(3) Ginhoux (4) Pollen

(5) Ting (6) Treutlin

Figure S1: The objective space visualization of EMDC on six real single-cell RNA-seq datasets.

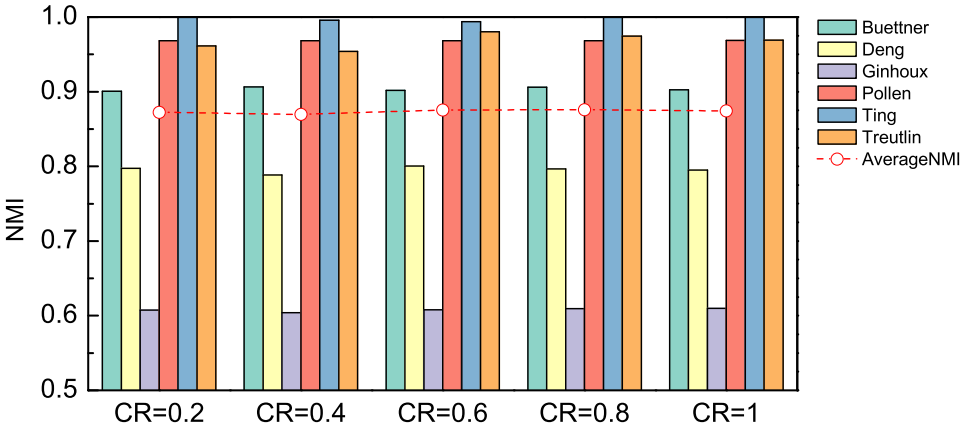


Figure S2: Performance comparisons of EMDC with varying crossover rates measured by *NMI* on six single-cell RNA-seq datasets.

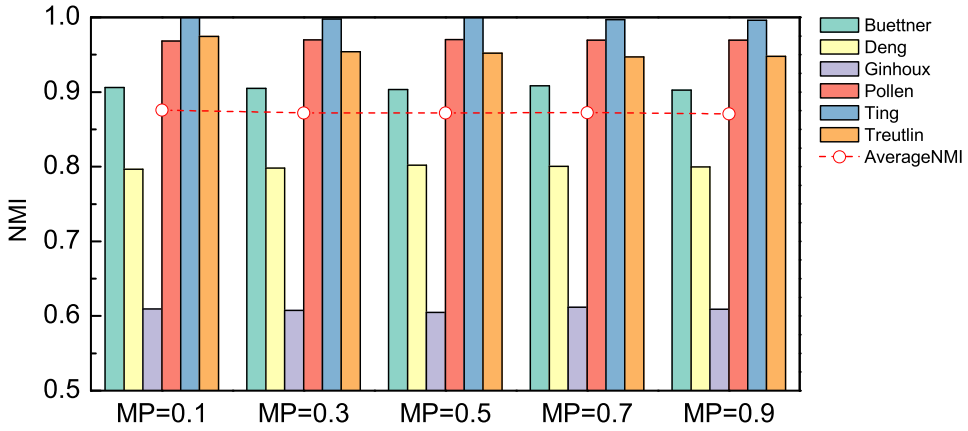


Figure S3: Performance comparisons of EMDC with varying mutation probabilities measured by *NMI* on six single-cell RNA-seq datasets.

Table S1: The detailed size of thirty synthetic single-cell RNA-seq datasets

DataSet No.	Cells	Genes	DataSet No.	Cells	Genes
1	100	2000	16	300	4000
2	100	2000	17	300	5000
3	100	3000	18	300	6000
4	100	3000	19	400	3000
5	100	4000	20	400	4000
6	100	4000	21	400	5000
7	100	5000	22	400	6000
8	100	5000	23	500	3000
9	100	6000	24	500	4000
10	100	6000	25	500	5000
11	200	3000	26	500	6000
12	200	4000	27	600	3000
13	200	5000	28	600	4000
14	200	6000	29	600	5000
15	300	3000	30	600	6000

Table S2: The variable table

Variable	Definition
Db	Davies-Bouldin Index
Dunn	Dunn Validity Index
Nc	the number of clusterings in the ensemble
FEs	the number of fitness function evaluations
n	the number of samples
m	the dimension of latent feature representation
N	the population size
$\Pi = \{\pi_1, \pi_2, ..., \pi_d\}$	the ensemble with d basic clusterings
$\Pi_s = \{\Pi_1, \Pi_2, ..., \Pi_i, ..., \Pi_N\}$	different ensembles
$s_i = \left\{s_i^1, s_i^2, ..., s_i^j, ..., s_i^d\right\}$	the binary mask vector for Π_i
L_i	the i-th clustering result
FC	the fold change
$avg(\cdot)$	the average value function
g	the gene of the single-cell RNA-seq data
t	the expression value of the genes in t-th group
$R = \{R_1, R_2, ..., R_j, ..., R_d\}$	different latent feature representations
$R_j = \left\{r_{j(1)}, r_{j(2)}, ..., r_{j(i)}, ..., r_{j(n)}\right\}$ $(j = \{1, 2, ..., d\})$	the j-th latent feature representation
$r_{j(i)} = \left\{r_{j(i)}^1, r_{j(i)}^2, ..., r_{j(i)}^k, ..., r_{j(i)}^m\right\}$	the i-th element in the j-th latent feature representation
$r_{j(i)}^k$	the k-th dimension of the i-th sample in the representation R_j
$X = \{x_1, x_2, ..., x_i, ..., x_n\}$	the gene expression matrix with n samples
x, y	two data samples
$x_i = \left\{x_i^1, x_i^2, ..., x_i^{\tilde{m}}\right\}$	each cell in X with $\tilde{m}$ genes
$\widehat{x}_i$	the reconstructed vector
$\varphi_\theta(\cdot)$	the encoder
$\phi_\theta(\cdot)$	the decoder
$\sigma(\cdot)$	the activation function
θ	the model parameter
l	the layer index
W,$\widehat{W}$	the weight matrices
b,$\widehat{b}$	the bias vectors
α	the non-zero gradient
$Loss(\cdot)$	the loss function
K	the number of clusters in the clustering
C_i	the i-th cluster in the clustering
c_i	the cluster centroid of the i-th cluster
$d(\cdot, \cdot)$	the Euclidean distance
$p_{max} = \left\{p_{max}^1, p_{max}^2, ..., p_{max}^d\right\}$	the upper bound of each individual
$p_{min} = \left\{p_{min}^1, p_{min}^2, ..., p_{min}^d\right\}$	the lower bound of each individual
p_{i_1},p_{i_2}	the parent individuals
v_k,v_i	the offspring individuals
η_c,η_m	the distribution index
μ_j,r_2	random numbers chosen from $[0, 1]$
CR	the crossover rate
MP	the mutation probability
K	the number of equality constraints
J	the number of inequality constraints
p^*	a Pareto-optimal solution
$\tilde{P}$	the mating pool
V	an offspring population
U	a new population
π_e	the obtained clustering
π_t	the ground truth clustering
n_e	the numbers of clusters in π_e
n_t	the numbers of clusters in π_t
n_e^i	the sample size in cluster i of π_e
n_t^j	the sample size in cluster j of π_t
n_{ij}	the intersection sample size between cluster i of π_e and cluster j of π_t
NMI	the Normalized Mutual Information
ARI	the Adjusted Rand Index

Table S3: The correspondence between the single-cell RNA-seq clustering problem and the multi-objective optimization

Variable	Definition in the single-cell RNA-seq clustering problem	Definition in the multiobjective optimization
d	the number of different low-dimensional latent feature representations	the number of variables in the individual
M	M=3	the number of objective functions
$F_i(\cdot)$	$F_1(\cdot) = \min(Db)$, $F_2(\cdot) = \max(Dunn)$, $F_3(\cdot) = \min(Nc)$	the i-th objective function
$F(\cdot)$	$F(\cdot) = \{\min(Db), \max(Dunn), \min(Nc)\}$	the objective functions
$P = \{p_1, p_2, ..., p_N\}$	N different ensembles	the population with N individuals
$p_i = \{p_i^1, p_i^2, ..., p_i^d\}$, $i = \{1, 2, ..., N\}$	an ensemble (a candidate solution)	the i-th individual in the population
$\overline{PS}$	multiple clustering solutions	the Pareto set

Table S4: Time complexity of different clustering methods, where I_k is the number of iterations for the convergence of k-means clustering, I_t is the number of iterations, I is the number of neurons in hidden layers, n is the number of samples, d is the number of dimension, K is the number of clusters, M is the ensemble size, N is the population size, m is the number of objective functions, T is the size of neighbors, r is the feature dimension of dataset after factorization, and t is the number of iterations in NMF.

Algorithm	LCE	KM	SC	SSC	ECC	CDP
Time Complexity	$O(n^2 \log n)$	$O(I_k K n d)$	$O(n^3)$	$O(I_t n^3)$	$O(I_t n K M)$	$O(n^2)$
Algorithm	SIMLR	MPSSC	NSGAIII	MOEA/D	EMEP	EMDC
Time Complexity	$O(n^2 d + n^3)$	$O(I_t n^3)$	$O(nI^2 + I_t N(n^2 d + n^3 + mN))$	$O(nI^2 + I_t N(n^2 d + n^3 + mT))$	$O(ndrt + I_t N(n^2 d + n^3 + mT))$	$O(nI^2 + I_t N(n^2 d + n^3 + mN))$

Table S5: Performance comparisons of different clustering algorithms on thirty synthetic single-cell RNA-seq datasets measured by *NMI*

DataSet No.	LCE	KM	SC	SSC	ECC	CDP	SIMLR	MPSSC	EMDC
1	0.1745	0.3724	0.0495	0.2559	0.4552	0.0421	0.3891	0.2452	**0.5776**
2	0.0631	0.3081	0.1841	0.2945	0.3577	0.0690	0.3357	0.2608	**0.5469**
3	0.1277	0.5734	0.0541	0.3882	0.5955	0.0551	0.4915	0.4727	**0.7255**
4	0.1206	0.5248	0.0741	0.5100	0.4959	0.0161	0.3088	0.5278	**0.7190**
5	0.4915	0.4211	0.2138	0.4309	0.5928	0.2962	0.5411	0.4998	**0.7264**
6	0.4416	0.4788	0.3906	0.4008	0.5187	0.1785	0.5636	0.3991	**0.6973**
7	0.3920	0.6312	0.3352	0.4013	0.6851	0.4158	0.4415	0.4728	**0.9014**
8	0.5972	0.5104	0.4567	0.4540	0.6394	0.1725	0.4067	0.5542	**0.8995**
9	0.4461	0.5543	0.3782	0.4872	0.9287	0.1836	0.5061	0.5764	**0.9825**
10	0.3636	0.6319	0.3701	0.4377	0.8190	0.2182	0.5757	0.4408	**0.9386**
11	0.2284	0.3466	0.0956	0.3339	0.3627	0.0478	0.3690	0.3411	**0.5353**
12	0.0478	0.4328	0.1399	0.3940	0.4571	0.0065	0.2808	0.3720	**0.5910**
13	0.5496	0.4903	0.2666	0.4485	0.4885	0.0216	0.4821	0.4797	**0.6706**
14	0.2469	0.6409	0.4149	0.4121	0.8275	0.0459	0.4492	0.3866	**0.8880**
15	0.2691	0.4590	0.0773	0.3738	0.6014	0.0248	0.4007	0.4102	**0.7617**
16	0.0385	0.6119	0.1323	0.4285	0.7787	0.0437	0.4584	0.4343	**0.8280**
17	0.0405	0.4592	0.0735	0.4266	0.5130	0.0514	0.3681	0.4857	**0.6450**
18	0.2211	0.4322	0.2314	0.3857	0.5368	0.0678	0.4561	0.4701	**0.5967**
19	0.3110	0.4281	0.1397	0.3805	**0.6634**	0.0569	0.5284	0.3915	0.5695
20	0.8284	0.5795	0.1990	0.4805	0.8014	0.0131	0.5182	0.4855	**0.9195**
21	0.0284	0.3412	0.0631	0.3615	0.4000	0.0430	0.2476	0.3647	**0.4358**
22	0.0353	0.4049	0.2220	0.3644	0.4578	0.0587	0.3752	0.4018	**0.5088**
23	0.4161	0.4786	0.3269	0.4808	0.5756	0.0775	0.4365	0.4706	**0.6359**
24	0.6016	0.5175	0.2685	0.4934	0.5597	0.0765	0.5112	0.4962	**0.6611**
25	0.7260	0.5625	0.1635	0.4995	0.8874	0.0551	0.6171	0.5114	**0.9280**
26	0.0594	0.3459	0.1436	0.3144	0.3856	0.0473	0.2615	0.3149	**0.4670**
27	0.0310	0.3735	0.0928	0.4065	0.4833	0.0147	0.2130	0.4193	**0.4978**
28	0.1450	0.4572	0.2741	0.4631	0.4990	0.0630	0.4365	0.4897	**0.5991**
29	0.5223	0.4908	0.3693	0.4138	**0.7405**	0.1150	0.4839	0.4466	0.6155
30	0.7596	0.5789	0.3207	0.5144	0.9164	0.1387	0.5308	0.5183	**0.9409**
Avg.	0.3108	0.4813	0.2174	0.4145	0.6008	0.0905	0.4328	0.4380	**0.7003**
Wilcoxon test	H_1	H_1	H_1	H_1	H_1	H_1	H_1	H_1	N/A

Table S6: Performance comparisons of different clustering algorithms on thirty synthetic single-cell RNA-seq datasets measured by *ARI*

DataSet No.	LCE	KM	SC	SSC	ECC	CDP	SIMLR	MPSSC	EMDC
1	0.0926	0.3744	0.0426	0.2518	0.4307	0.0115	0.4150	0.2328	**0.5468**
2	-0.0029	0.3050	0.1930	0.2650	0.3753	0.0385	0.2988	0.2656	**0.5960**
3	0.0531	0.4877	0.0120	0.3158	0.5502	0.0348	0.4221	0.4026	**0.5947**
4	0.0362	0.5406	0.0461	0.5202	0.5071	-0.0251	0.2967	0.5150	**0.6634**
5	0.4502	0.4016	0.2157	0.3980	0.5785	0.2448	0.5460	0.5050	**0.7174**
6	0.3567	0.4140	0.3410	0.3830	0.5226	0.0793	0.5937	0.3845	**0.6142**
7	0.2924	0.6138	0.3710	0.3664	0.6877	0.4317	0.3639	0.4985	**0.9245**
8	0.5198	0.4540	0.3434	0.3827	0.6215	0.0903	0.3042	0.5545	**0.9199**
9	0.3206	0.5542	0.4362	0.4947	0.9533	0.1351	0.5315	0.5431	**0.9885**
10	0.2695	0.6374	0.2940	0.4150	0.8653	0.2163	0.6169	0.4101	**0.9568**
11	0.1384	0.2945	0.0601	0.2921	0.3078	0.0225	0.3476	0.3070	**0.4079**
12	0.0099	0.3667	0.0080	0.3495	0.3991	0.0057	0.2734	0.3561	**0.5980**
13	0.4407	0.4136	0.3227	0.3858	0.4077	0.0116	0.4214	0.4202	**0.5843**
14	0.2110	0.6434	0.3592	0.3682	0.8798	0.0159	0.4084	0.3828	**0.9180**
15	0.2271	0.4270	0.1193	0.4119	0.6686	0.0454	0.3831	0.3965	**0.8244**
16	0.0012	0.5652	0.0223	0.3577	0.7974	0.0538	0.3934	0.3723	**0.8556**
17	0.0026	0.3769	0.0565	0.3590	0.4339	-0.0166	0.2648	0.4237	**0.5508**
18	0.1684	0.3667	0.2152	0.3259	0.4645	-0.0334	0.4257	0.4257	**0.5553**
19	0.2390	0.3518	0.0862	0.3256	**0.5796**	0.0298	0.4485	0.3459	0.5185
20	0.7634	0.5019	0.0739	0.4636	0.7971	-0.0272	0.4577	0.4343	**0.9416**
21	0.0005	0.2616	0.0478	0.3083	0.3338	0.0392	0.2363	0.3182	**0.3627**
22	0.0049	0.3443	0.2642	0.3246	0.4002	0.0201	0.3606	0.3691	**0.4736**
23	0.3421	0.4042	0.3509	0.4244	0.4950	0.0618	0.4002	0.4277	**0.5473**
24	0.4819	0.4254	0.1590	0.4185	0.4602	0.0456	0.4321	0.4285	**0.5502**
25	0.6509	0.4788	0.0528	0.4332	0.8768	0.0458	0.6030	0.4444	**0.9547**
26	0.0052	0.2660	0.1279	0.2895	0.3260	0.0443	0.2300	0.2962	**0.3530**
27	0.0031	0.2874	0.1005	0.3617	0.4152	-0.0286	0.1693	0.3746	**0.4407**
28	0.1047	0.3677	0.2049	0.3805	0.4218	0.0517	0.3808	0.4266	**0.4987**
29	0.4190	0.4158	0.4055	0.3592	**0.6837**	0.1705	0.4018	0.3986	0.5281
30	0.6391	0.5040	0.2284	0.4653	0.9293	0.1411	0.4533	0.4605	**0.9648**
Avg.	0.2414	0.4282	0.1853	0.3732	0.5723	0.0652	0.3960	0.4040	**0.6650**
Wilcoxon test	H_1	H_1	H_1	H_1	H_1	H_1	H_1	H_1	N/A

Table S7: Performance comparisons of different clustering algorithms on six single-cell RNA-seq datasets measured by *NMI*

DataSet	LCE	KM	SC	SSC	ECC	CDP	SIMLR	MPSSC	EMDC
Buettner	0.4818	0.4297	0.8026	0.7638	0.4609	0.6137	0.8381	0.8342	**0.9060**
Deng	0.7216	0.7458	0.6782	0.6579	0.7430	0.3644	0.7515	0.7554	**0.7966**
Ginhoux	0.4313	0.3758	0.5840	0.5828	0.4301	0.0232	0.3878	0.6636	**0.6094**
Pollen	0.9505	0.8107	0.8948	0.9446	0.9067	0.4239	0.9481	0.9360	**0.9684**
Ting	0.8799	0.5128	0.8622	0.9755	0.8101	0.5363	0.9754	0.9755	**1.0000**
Treutlin	0.7998	0.7800	0.7226	0.7022	0.6061	0.3847	0.6881	0.8010	**0.9746**
Avg.	0.7108	0.6091	0.7574	0.7711	0.6595	0.3910	0.7648	0.8277	**0.8758**
Wilcoxon test	H_1	H_1	H_1	H_1	H_1	H_1	H_1	H_0	N/A

Table S8: Performance comparisons of different clustering algorithms on six single-cell RNA-seq datasets measured by *ARI*

DataSet	LCE	KM	SC	SCC	ECC	CDP	SIMLR	MPSSC	EMDC
Buettner	0.3179	0.3690	0.8169	0.7631	0.4472	0.4977	0.8586	0.8443	**0.9318**
Deng	0.5004	0.4807	0.5005	0.3804	0.5807	0.0870	0.4772	0.4783	**0.6474**
Ginhoux	0.3930	0.2814	0.6095	0.5937	0.3849	0.0050	0.3239	0.7317	**0.6428**
Pollen	0.9429	0.5581	0.8044	0.9292	0.8874	0.0746	0.9408	0.9328	**0.9563**
Ting	0.7769	0.3441	0.7974	0.9784	0.6783	0.4073	0.9803	0.9784	**1.0000**
Treutlin	0.8353	0.8317	0.5482	0.5242	0.4534	0.1625	0.5114	0.6117	**0.9808**
Avg.	0.6277	0.4775	0.6795	0.6948	0.5720	0.2057	0.6820	0.7629	**0.8599**
Wilcoxon test	H_1	H_1	H_1	H_1	H_1	H_1	H_1	H_0	N/A

Table S9: Performance comparisons of different multiobjective optimization algorithms on six single-cell RNA-seq datasets measured by *NMI*

DataSet	NSGAIII	MOEA/D	EMEP	EMDC
Buettner	0.8999	0.7873	0.8629	**0.9060**
Deng	0.7797	0.7477	**0.8228**	0.7966
Ginhoux	0.6052	0.5571	0.5508	**0.6094**
Pollen	0.9673	0.9494	0.9674	**0.9684**
Ting	0.9978	0.9466	1.0000	**1.0000**
Treutlin	0.9582	0.8441	0.9310	**0.9746**
Average	0.8680	0.8054	0.8558	**0.8758**

Table S10: Performance comparisons of different multiobjective optimization algorithms on six single-cell RNA-seq datasets measured by *ARI*

DataSet	NSGAIII	MOEA/D	EMEP	EMDC
Buettner	0.9276	0.8261	0.9016	**0.9318**
Deng	0.5891	0.5305	**0.7202**	0.6474
Ginhoux	0.6347	0.6090	0.5325	**0.6428**
Pollen	0.9539	0.9002	0.9540	**0.9563**
Ting	0.9977	0.9360	1.0000	**1.0000**
Treutlin	0.9675	0.8256	0.9524	**0.9808**
Average	0.8451	0.7712	0.8434	**0.8599**

Table S11: Performance comparisons of EMDC with different dimension reduction methods on six single-cell RNA-seq datasets measured by *NMI*

DataSet	EMDC$_{NMF}$	EMDC$_{ICA}$	EMDC$_{PCA}$	EMDC$_{LPP}$	EMDC
Buettner	0.6008	0.5405	0.6498	0.1361	**0.9060**
Deng	0.7756	0.5654	**0.7975**	0.6634	0.7966
Ginhoux	0.4581	0.2131	0.4936	0.1076	**0.6094**
Pollen	0.9565	0.8905	**0.9718**	0.9630	0.9684
Ting	0.8995	0.8311	0.9572	0.8311	**1.0000**
Treutlin	0.5743	0.3024	0.8121	0.7381	**0.9746**
Avg.	0.7108	0.5572	0.7803	0.5732	**0.8758**

Table S12: Performance comparisons of EMDC with different dimension reduction methods on six single-cell RNA-seq datasets measured by *ARI*

DataSet	EMDC$_{NMF}$	EMDC$_{ICA}$	EMDC$_{PCA}$	EMDC$_{LPP}$	EMDC
Buettner	0.5415	0.4916	0.6638	0.0157	**0.9318**
Deng	0.5322	0.3231	0.6011	0.4100	**0.6474**
Ginhoux	0.4373	0.1883	0.5037	-0.0019	**0.6428**
Pollen	0.9327	0.8185	**0.9640**	0.9483	0.9563
Ting	0.8273	0.7658	0.9484	0.7658	**1.0000**
Treutlin	0.4989	0.1403	0.7957	0.6886	**0.9808**
Avg.	0.6283	0.4546	0.7461	0.4711	**0.8599**