nature machine intelligence

ARTICLES

https://doi.org/10.1038/s42256-021-00333-y

Simultaneous deep generative modelling and clustering of single-cell genomic data

Qiao Liu[1,2], Shengquan Chen[1], Rui Jiang[1]✉ and Wing Hung Wong[2,3]✉

Recent advances in single-cell technologies, including single-cell ATAC-seq (scATAC-seq), have enabled large-scale profiling of the chromatin accessibility landscape at the single-cell level. However, the characteristics of scATAC-seq data, including high sparsity and high dimensionality, have greatly complicated the computational analysis. Here, we propose scDEC, a computational tool for scATAC-seq analysis with deep generative neural networks. scDEC is built on a pair of generative adversarial networks, and is capable of simultaneously learning the latent representation and inferring cell labels. In a series of experiments, scDEC demonstrates superior performance over other tools in scATAC-seq analysis across multiple datasets and experimental settings. In downstream applications, we demonstrate that the generative power of scDEC helps to infer the trajectory and intermediate state of cells during differentiation and the latent features learned by scDEC can potentially reveal both biological cell types and within-cell-type variations. We also show that it is possible to extend scDEC for the integrative analysis of multi-modal single cell data.

The organization of chromatin accessibility across the whole genome reflects an epigenetic landscape of gene regulation[1,2]. With the recent development in single-cell technology, it becomes feasible to characterize the epigenetic landscape of individual cells[3]. In particular, scATAC-seq is an efficient method for the study of variation in chromatin accessibility both between and within populations at the single-cell level[4,5]. However, the analysis of scATAC-seq presents unique methodological challenges due to the high dimensionality (hundreds of thousands possible peaks) and high data sparsity (only 1–10% peaks are detected per cell)[6].

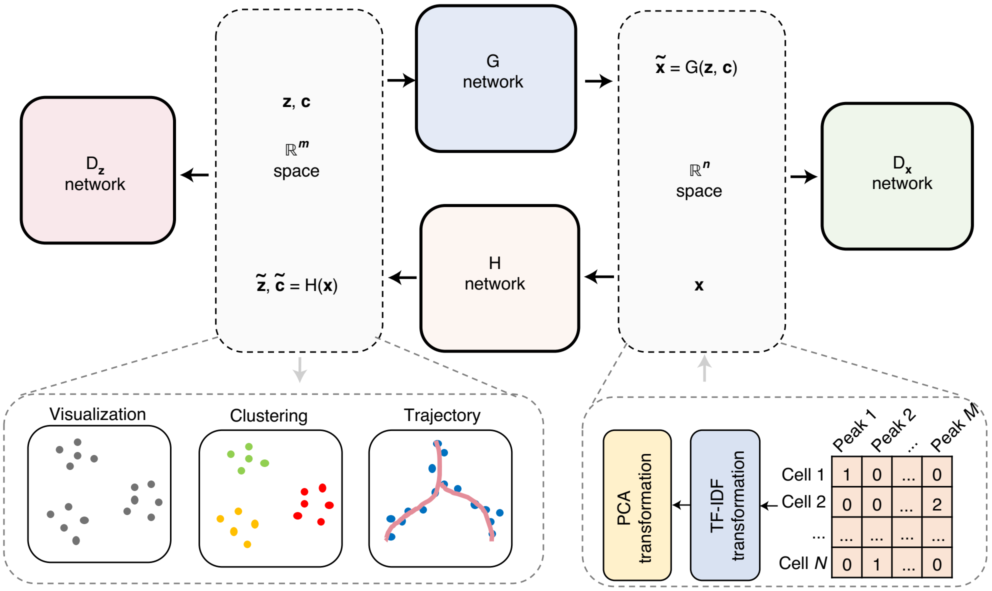
Several computational approaches have been proposed to tackle the challenges in scATAC-seq analysis. scABC estimated weights of cells based on the number of distinct reads and applied a weighted K-medoids clustering to infer cell types[7]. cisTopic applied latent Dirichlet allocation (LDA) as a probabilistic model to identify the *cis*-regulatory topics enriched in different cells by simultaneously optimizing topic–cell probability and region–topic probability[8]. Cusanovich et al. proposed a pipeline that performs the term frequency–inverse document frequency transformation (TF-IDF) and singular value decomposition (SVD) iteratively to get a low-dimensional representation of scATAC-seq data[4,9]. Scasat introduced another pipeline which involved Jaccard similarity measure and multidimensional scaling (MDS) to reduce the high dimensionality in scATAC data[10]. SnapATAC divided the genome into bins of equal size, built a bins-by-cells binary count matrix and then applied principle component analysis (PCA) for a dimension reduction[11]. Recently, deep generative models have emerged as a powerful framework for both representation learning and data generation[12,13,14]. A newly developed method, SCALE, utilized a variational autoencoder (VAE) to learn the latent features of scATAC-seq data and then used a K-means by default for clustering the latent features[15].

Here, we propose a new approach for analysing scATAC-seq data by simultaneously learning the deep embedding and clustering of the cells in an unsupervised manner. Our method, named scDEC, is based on learning a pair of generative adversarial networks (GANs; Fig. 1). Such a symmetrical and paired GAN architecture has recently been successfully applied to image style transfer[16] and density estimation[17] while we adopt this architecture to the new task of unsupervised clustering and apply it to the analysis of single-cell genomic data. Unlike the methods discussed above, where an external method (for example, K-means) is typically required for clustering the latent features, our model directly models the cell clustering process. Thus, cell clustering and latent feature representation learning are jointly optimized during the training process. In other words, scDEC enables simultaneous learning of latent features and cell clustering. We demonstrate the advantage of this approach in a series of experiments, where scDEC shows superiority over competing methods. We also illustrate several downstream applications of scDEC in scATAC-seq analysis, including trajectory inference, donor effect removal and latent feature interpretation. Finally, we extend scDEC to multi-modal single-cell analysis and demonstrate its effectiveness in a real data example.

Results

Overview of scDEC model. scDEC consists of two GAN models, which are utilized for transformations between latent space and data space (Fig. 1). The scATAC-seq data are first preprocessed through a TF-IDF transformation and a PCA dimension reduction before being fed to the scDEC model. Assuming that the input scATAC-seq data contains K cell types, then a continuous latent variable $\mathbf{z}$ and a discrete latent variable $\mathbf{c}$ are introduced, where $\mathbf{z} \sim \mathcal{N}(0, I)$ and $\mathbf{c} \sim \mathrm{Cat}(K, \mathbf{w})$, respectively. We also provide an approach for estimating the number of cell subpopulations if K is unknown (Methods). The forward transformation through the G network can be considered as a process of conditional generation given an encoded style ($\mathbf{z}$) and an indicated cluster label ($\mathbf{c}$). The backward transformation through the H network aims at encoding a data point $\mathbf{x}$ to the latent

[1]Ministry of Education Key Laboratory of Bioinformatics, Research Department of Bioinformatics at the Beijing National Research Center for Information Science and Technology, Center for Synthetic and Systems Biology, Department of Automation, Tsinghua University, Beijing, China. [2]Department of Statistics, Stanford University, Stanford, CA, USA. [3]Department of Biomedical Data Science, Bio-X Program, Center for Personal Dynamic Regulomes, Stanford University, Stanford, CA, USA. ✉e-mail: ruijiang@tsinghua.edu.cn; whwong@stanford.edu



	Peak 1	Peak 2	...	Peak *M*
Cell 1	1	0	...	0
Cell 2	0	0	...	2
...	...	...	...	...
Cell *N*	0	1	...	0

Fig. 1 | The illustration of scDEC model. The read count matrix of scATAC-seq will first be preprocessed by a TF-IDF transformation and a PCA dimension reduction (for example, $n = 20$) before it is fed to the scDEC model. In the latent space, latent variables **z** and **c**, sampled from a Gaussian distribution and a Category distribution respectively, will be concatenated together before they are fed to the G network. The H network has two outputs of which one corresponds to the latent embedding ($\tilde{\mathbf{z}}$) and one corresponds to the estimated cluster label ($\tilde{\mathbf{c}}$) through a softmax function. The D_x network works as a discriminator for discerning the true scATAC-seq data (**x**) from the generated data ($\tilde{\mathbf{x}}$). The D_z network is another discriminator for distinguishing the learned continuous latent variable ($\tilde{\mathbf{z}}$) from the real continuous latent variable (**z**).

space and inferring the cluster label, simultaneously. If we assume the last layer of the H network contains m nodes ($m > K$), then $\tilde{\mathbf{z}}$ denotes the output of the first m-K nodes and $\tilde{\mathbf{c}}$ denotes the output of the remaining K nodes with an additional softmax function. D_x and D_z are two discriminator networks that are used for matching the distributions of data $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{z}}$ to the empirical distribution of the data and latent variable distribution, respectively. (G, D_x) and (H, D_z) can be considered as two GAN models that are jointly trained. The G and H network each contain 10 fully connected layers while D_x and D_z have two fully connected layers each (see detailed hyperparameters in Supplementary Table 1). Note that the weight **w** in the category distribution is also learned automatically via an updating scheme according to the feedback of inferred cluster labels by $\tilde{\mathbf{c}}$ (Methods). After model training, the cluster labels are inferred based on $\tilde{\mathbf{c}}$ (Methods). The output of the last layer of the H network combined with $\tilde{\mathbf{z}}$ and $\tilde{\mathbf{c}}$ (before softmax) is used as a low-dimensional representation for downstream analysis such as data visualization and trajectory analysis.

scDEC automatically identifies cell types in scATAC-seq data. To demonstrate the ability of scDEC to reveal differences between cell subpopulations and identify cell types in an unsupervised manner, we tested scDEC on four benchmark scATAC-seq datasets across different numbers of cells and cell types (see statistics and abbreviations in Supplementary Fig. 1). Specifically, scDEC was benchmarked against six baseline methods, including scABC[7], SCALE[15], cisTopic[8], Cusanovich2018[4,9], Scasat[10] and SnapATAC[11] (Methods). The performance of a method was evaluated on (1) whether different cell subpopulations can be clearly separated in a low-dimensional space and (2) whether true cell type labels can be accurately inferred by clustering. To address the first question, we applied each method to conduct a dimension reduction or to extract the latent features. The latent dimension was set to 15 for the two datasets with relatively smaller numbers of cells and cell types, and 20 for the two larger datasets. For each method, we constructed a t-distributed stochastic neighbour embedding (t-SNE)[18] or uniform manifold approximation and projection (UMAP)[19] plot based on the latent features and then visualized with fluorescence-activated cell sorting (FACS) cell labels on the plot to see whether the subpopulations were well separated. To address the second question, we evaluated the clustering results of each method based on FACS sorting cell labels using three commonly used metrics, namely normalized mutual information (NMI), adjusted Rand index (ARI) and homogeneity score (Methods). Since five of the methods (except scABC) focused on learning a low-dimensional representation and require an additional clustering step, we used Louvain clustering[20], which was recommended by a benchmark study[6], for clustering the latent features learned by these methods. The results are summarized below for each dataset.

InSilico dataset. This dataset[5] is an in silico mixture constructed by artificially combining six individual scATAC-seq experiments, which were separately conducted on a different cell line. It is observed that cells from a minor cell type TF-1 (6.83%, in purple) are dispersed into several clusters by SCALE, Cusanovich2018, Scasat and SnapATAC while cisTopic and scDEC can maintain the close distance in the low-dimensional representation (Fig. 2a). scDEC achieves an NMI of 0.871, an ARI of 0.896 and a homogeneity of 0.866, which outperforms the best baseline method scABC (NMI = 0.822, AIR = 0.855 and homogeneity = 0.840) by a noticeable margin (Fig. 2e and Supplementary Fig. 2).

Forebrain dataset. This dataset[21] was derived from P56 mouse forebrain cells which contained eight different cell groups in adult mouse forebrain (EX: excitatory neuron, IN: inhibitory neuron, AC: astrocyte, OG: oligodendrocyte, MG: microglia). Interestingly, all the baseline methods fail to distinguish three subtypes of excitatory neuron cells (EX1, EX2 and EX3) while scDEC shows a relatively clear separation among these three subpopulations of cells (Fig. 2b). Again, scDEC demonstrates a superior clustering performance by

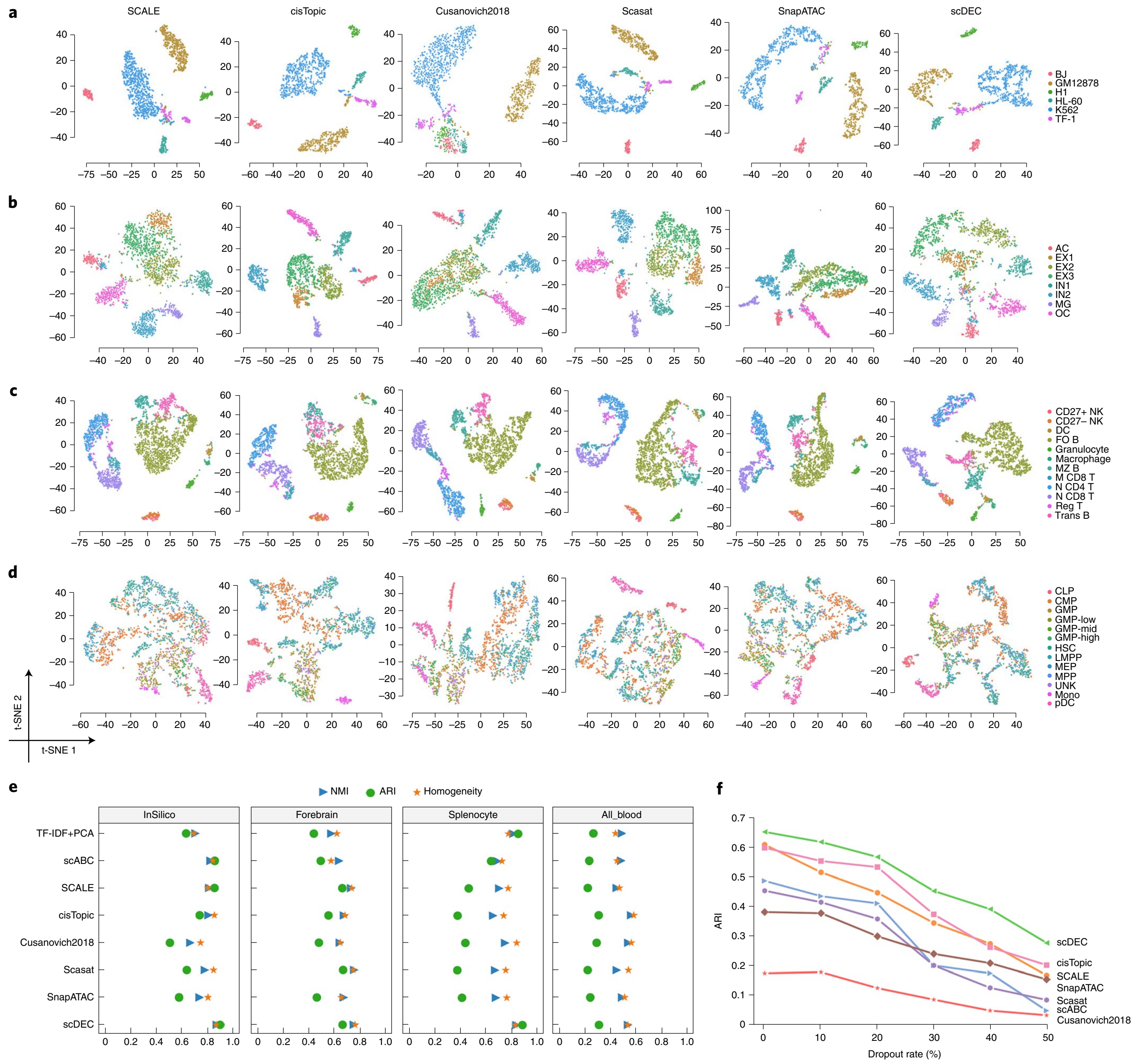


Fig. 2 | Evaluation of scDEC compared with other baseline methods. a–d, Visualization of the InSilico (**a**), Forebrain (**b**), Splenocyte (**c**) and All blood (**d**) datasets by different methods. **e**, Clustering results of the different methods across the four datasets. **f**, Performance of different methods under different dropout rates on the Forebrain dataset.

achieving the highest NMI of 0.750, ARI of 0.663 and homogeneity of 0.759 (Fig. 2e and Supplementary Fig. 3).

Splenocyte dataset. This dataset[22] was collected from a mixture of mouse splenocytes after removing red blood cells, which finally resulted in 12 cell subpopulations (CD27+ NK: CD27+ natural killer cell, CD27- NK: CD27− natural killer cell, DC: dendritic cell, Fo B: follicular B cell, MZ B: marginal zone B cell, M CD8 T: memory CD8 T cell, N CD4 T: naïve CD4 T cell, N CD8 T: naïve CD8 T cell, Reg T: regulatory T cell, Trans B: transitional B cell). A major cell type, follicular B cells (FO B, 42.89%), together with marginal zone B cells (MZ B) and transitional B cells (Trans B) are more or less mixed together by all baseline methods while scDEC illustrates a clearer separation (Fig. 2c). As the largest dataset (around 3,000 cells) among the four, scDEC still achieves the highest NMI of 0.839, ARI of 0.884 and homogeneity of 0.829 (Fig. 2e and Supplementary Fig. 4).

All blood dataset. This dataset[23] involves cellular differentiation of multipotent cells during human haematopoiesis, containing 13 cell subpopulations in total (CLP: common lymphoid progenitor, CMP: common myeloid progenitor, GMP: granulocyte-macrophage progenitor, HSC: hematopoietic stem cell, LMPP: lymphoid- primed multipotent progenitor, MEP: megakaryocyte-erythrocyte progenitor, MPP: multipotent progenitors, UNK: human natural killer, Mono: monocyte, pDC: plasmacytoid dendritic cell). Three types of cells, including Mono, pDC and CLP cells, can only be separated from other cells by cisTopic, Scasat and scDEC (Fig. 2d). scDEC still achieves the highest ARI (0.309) among all compared methods. The overall clustering performance is comparable with Cusanovich2018 and slightly lower than cisTopic (Fig. 2e and Supplementary Fig. 5).

scDEC achieves the best or second best (in one case) clustering results across multiple scATAC-seq datasets. scDEC shows consistently superior performance if we replace the Louvain clustering with the commonly used K-means clustering for the compared methods (Supplementary Fig. 6). The t-SNE visualizations of scDEC coloured by the cluster label identified by scDEC across the above four benchmark datasets are also provided (Supplementary Fig. 7). We also note that the performance of scDEC is not sensitive to the dimension of latent features (Supplementary Fig. 8).

Next, we further investigate the performance of different methods at different dropout rates, in order to assess the ability of handing scATAC-seq data with different degree of sparsity. We downsampled the original reads in the Forebrain dataset by randomly dropping out the non-zero entities in the read count matrix with probability equal to the dropout rate. scDEC consistently demonstrates the best performance with respect to the ARI metric for clustering at different dropout rates ranging from 0 to 50%. At the dropout rate of 50%, scDEC achieves an ARI of 0.279, compared with 0.202 of the best method cisTopic (Fig. 2f).

scDEC facilitates cell-type-specific motif discovery and trajectory inference. We next explored whether scDEC can help identity cell-type specific motifs, which is essential for understanding the context-specific gene regulation. To achieve this, we first applied the scDEC model to the Forebrain dataset[21] to infer the cluster label for each individual cell, and used chromVAR[24] to identify cluster-specific enriched motifs from the JASPAR database[25]. We ranked cluster-specific enriched motifs (Methods) and discovered several significant motif enrichment patterns (Fig. 3a, Supplementary Table 2). Both single cluster-specific motifs and the co-occurrence of motifs in two (cluster 1 and 6) or three clusters (cluster 2, 3 and 4) are observed, which might reveal the co-regulation mechanism underlying the corresponding multiple TFs. For example, *En1*, which is enriched in cluster 1 (one-sided Mann–Whitney U test, $p=6.14\times10^{-51}$), is a well-known marker for the brain fate in astrocytes (AC)[26]. It is reported that *Neurod2* ($p=4.50\times10^{-239}$) regulates the cortical projection neuron, which constitutes the major excitatory neuron (EX) population[27]. *Meis1* (p-value $=6.68\times10^{-59}$) was known to have crucial functions in neural differentiation from neural progenitors[28]. *Vax1* ($p=2.84\times10^{-126}$) is a novel homeobox-containing gene that regulates the development of the basal forebrain[29]. The impact of *Elk1* ($p=1.87\times10^{-71}$) deficiency was proved to indicate the microglial (MG) activation[30]. The compound loss of *Sox9* ($p=3.81\times10^{-137}$) may lead to a further decrease in oligodendrocyte (OC) progenitors[31]. Interestingly, among the three similar cell types (EX1–EX3), we also discovered several motifs that were only enriched in one or two specific clusters that correspond to EX cells identified by scDEC (Supplementary Fig. 9). Several example literature-validated motifs are demonstrated in the t-SNE visualization according to the enrichment score calculated by chromVAR (Fig. 3b).

Next, we applied scDEC to trajectory inference during the haematopoiesis differentiation. We collected the cells from the donor BM0828 of the All blood dataset, which contains 533 cells across 7 subpopulations at different stages of differentiation. After obtaining the low-dimensional representation and the inferred cluster labels of scATAC-seq data, the smooth curves are annotated with cell lineages using Slingshot software[32] (Fig. 3c). The smooth curves with a tree-based structure are largely consistent with the true haematopoietic differentiation tree. Although it has been proved that common myeloid progenitors (CMPs) can differentiate into both granulocyte-macrophage progenitors (GMPs) and megakaryocyte-erythrocyte progenitors (MEPs)[33], only the differentiation path from CMP to MEP is observed in this dataset. We then took the cells from multipotent progenitors (MPPs), lymphoid- primed multipotent progenitors (LMPPs) and common lymphoid progenitors (CLPs) for a further study, where there exists a differentiation path (MPP→LMPP→CLP). To fully exploit the generation power of scDEC, we first left LMPP out as the target cells for imputation and trained scDEC based on the remaining cells comprising only MPP and CLP cells. Then we imputed data by interpolating the latent label indicator (Methods) and visualized the imputed data together with the true data. Interestingly, when the interpolation coefficient α changes from 0 to 1, the imputed data seem to capture the dynamics differentiation path from MPP to CLP. Specifically, the generated scATAC-seq data are similar to the real LMPP data according to t-SNE visualization when $\alpha=0.5$ (Fig. 3d). Next, we asked whether the interpolation on the latent indicator is a more effective way of data generation than directly interpolating on the raw scATAC-seq. We averaged all the scATAC-seq data of LMPP cells as a meta-cell and calculated the Pearson correlation between generated data and meta-cell. The data generated by scDEC achieve a notably higher correlation than data generated by direct interpolation and interpolation on PCA-reduced data (Fig. 3e and Supplementary Table 3). In summary, the generation power of scDEC sheds light on recovering the missing cell types of scATAC data and exploring the intermediate state of two neighbouring cell types of scATAC-seq data.

scDEC disentangles donor effect and promotes interpretation of latent features. Single-cell experiments are often conducted with notable differences in capturing time, equipment and even technology platforms, which may introduce batch effects to the data. To evaluate whether scDEC can automatically correct or alleviate batch effects in the training process, we collected three types (CLP, LMPP and MPP) of human haematopoietic cells from two donors with ID BM0828 (donor 1) and BM1077 (donor 2)[23]. We mixed the cells from the two donors together (200 cells from donor 1 and 180 cells from donor 2) and evaluated how well the variation due to cell type and donor was resolved in the embedding (that is, latent representation) learned by scDEC and alternative methods. Note that the latent dimension of each method was fixed to 13 and no donor information was revealed to each method. Since the embedding by scDEC depends on the number of clusters K, we varied K from 2 to 6 and examined the gap statistic plot (Fig. 4d), which exhibited two peaks at $K=3$ and $K=5$, respectively. The embedding results for scDEC and alternative methods are shown in Fig. 4a and Supplementary Figs. 10–13. The three cell types as well as the donor effects in two of the cell types are well captured by scDEC ($K=5$), cisTopics and SnapATAC, but not by SCALE, whereas the donor effect in the third cell type (CLP) is too small to be discernible. It is interesting that at $K=3$ (the first peak of the gap statistic) the clustering results by scDEC matches the three cell types almost perfectly. Specifically, SCALE is basically unable to clearly separate the three types of cell. cisTopic and SnapATAC cannot alleviate the donor effect in LMPP or MPP cells, as the same types of cell from two different donors were separated with a notable distance in the t-SNE plot (Fig. 4a). Considering the first mode where $K=3$, only 9 cells from donor 1 and 17 cells from donor 2 were wrongly clustered by scDEC, which illustrates a total error rate of 6.86%. Besides, scDEC also demonstrates an NMI of 0.754, ARI of 0.805 and homogeneity of 0.757, which outperforms other compared methods by a large margin (Fig. 4b and Supplementary Fig. 13). In this sense our method can be used to adjust for donor or batch effects in clustering and visualization.

Next, we carefully analysed the latent feature learned by scDEC by visualization. We noticed that features corresponding to the

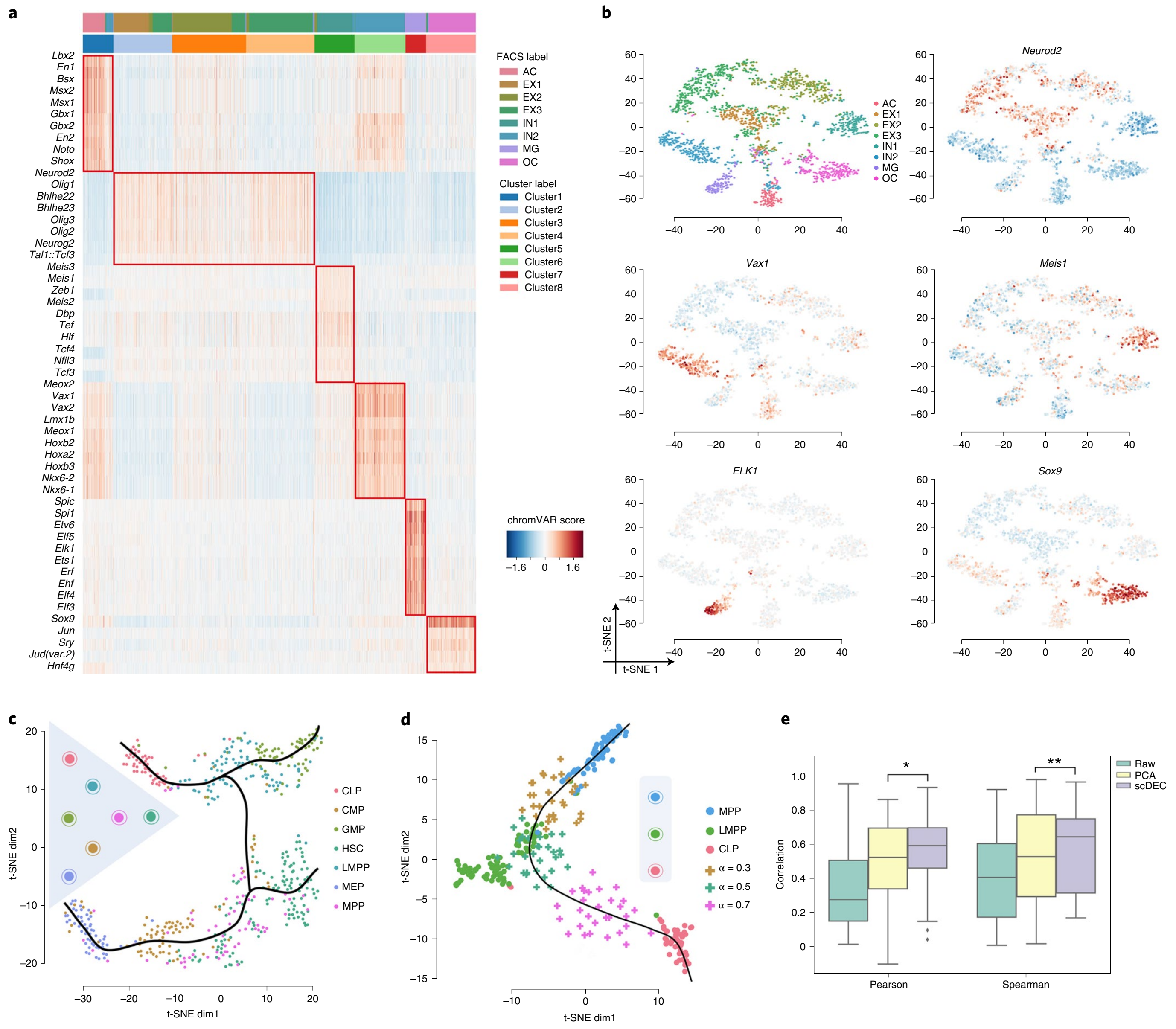


Fig. 3 | Cluster-specific motif recovery and trajectory inference. **a**, Heatmap of enriched motifs; each row denotes a motif and each column denotes a cell. Both cluster label and FACS label were provided and aligned. **b**, t-SNE visualization of several literature-validated motifs. **c**, Haematopoiesis differentiation trajectory inferred by scDEC. **d**, The generated intermediate state between MPP and CLP. Thirty data points were generated at different generation coefficients α. **e**, The generated intermediate scATAC data by interpolation on the latent label indicator has a higher correlation with the meta cell (the average profile of ground truth cells) than the scATAC-seq that were directly interpolated on the raw data and PCA reduced data in a box plot. (The box denotes the interquartile range (IQR) and the whiskers denote 1.5IQR beyond the low and high quartiles, $^{*}p<1.28\times10^{-16}$, $^{**}p<4.40\times10^{-8}$).

latent discrete variable (features 11–13) were highly correlated with biological cell type while other features more or less revealed within-cell-type variations (Fig. 4e). For example, feature 1 is highly expressed in the LMPP of donor 2 and the MPP of donor 1. Feature 10 can be a donor-specific indicator of LMPP. We proposed a strategy for mining motif information underlying the latent features (Supplementary Fig. 14). Through the strategy, the top-ranked motif ($p=1\times10^{-90}$) for feature 2 is SP1, which was proven to affect multiple haematopoietic lineages[34]. To sum up, the interpretable features in the latent space reveal both biological cell types and within-cell-type variations.

scDEC is capable of analysing large scATAC-seq data. We further examine whether scDEC is applicable to extremely large scATAC-seq datasets. We collected a dataset from a mouse atlas study, which contains 81,173 single cells from 13 adult mouse tissues using sci-ATAC-seq[9]. The original atlas study applies a computational pipeline to infer 40 cell types, which were regarded as 'reference' cell labels for the comparison of scDEC and other baseline methods. To investigate the scalability of scDEC, we randomly down-sampled the original dataset to a different scale of dataset and scDEC shows a consistently good agreement with the reference cell label (Fig. 4f). For the full scale of the dataset, scDEC achieves an NMI of 0.732, ARI of 0.614 and homogeneity of 0.693 while most methods failed to handle the full dataset due to the memory limitation (500 GB for the computational environment). We compared scDEC to the deep learning method SCALE and noticed that scDEC achieves a higher consistency with 'reference' label but a little slower

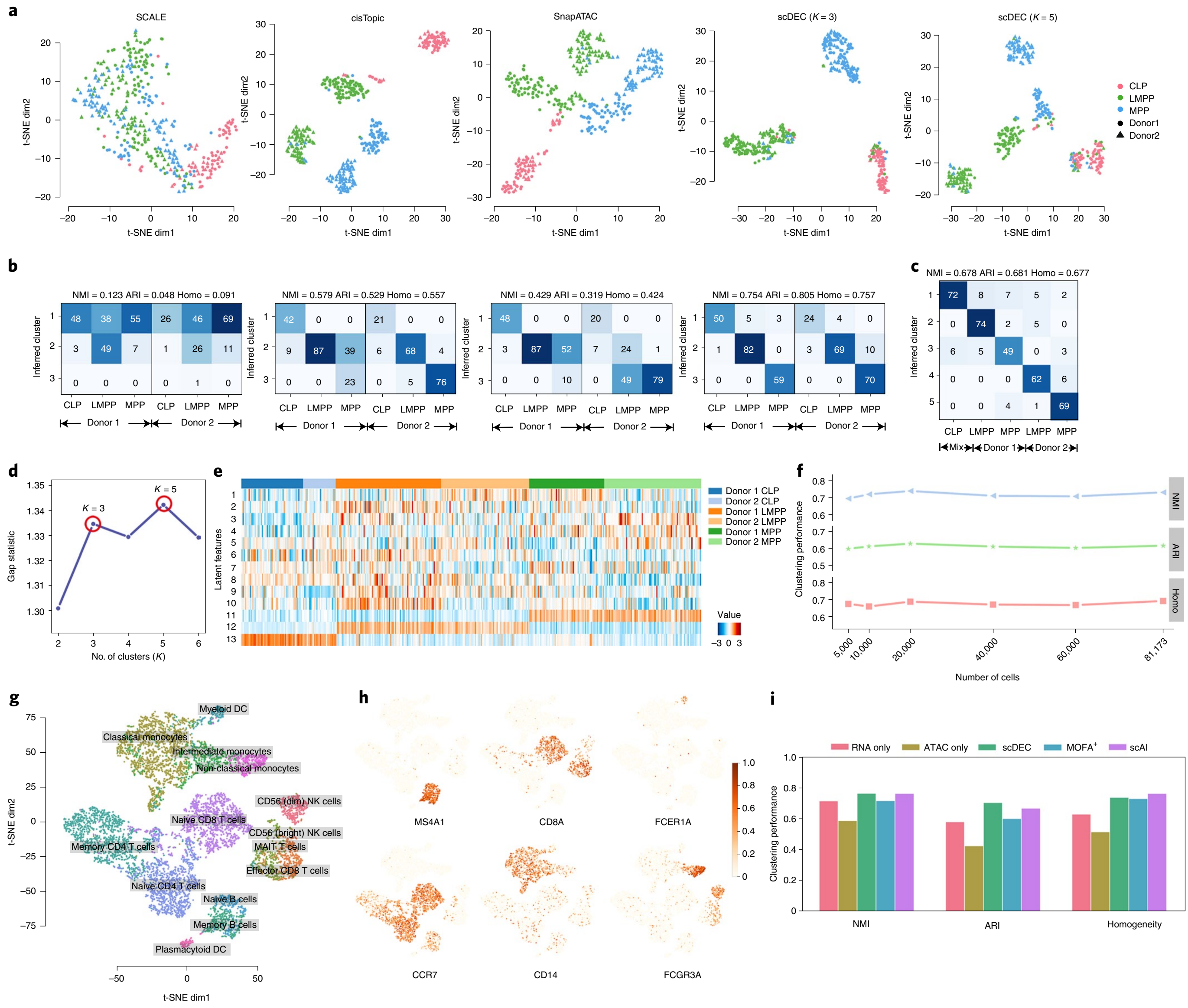


Fig. 4 | scDEC alleviates donor effect and is applicable to large dataset and multi-modal single cell dataset. a, The t-SNE visualization, for CLP, LMPP, MPP cells of the latent features learned by different methods. Colours denote different cell types and shapes (circle or triangle) represent which donor it comes from. For scDEC, different K (3 and 5) results in different latent-feature visualization. **b**, The confusion matrix of the clustering by scDEC and compared methods ($K=3$). The NMI, ARI and homogeneity are also annotated on the top of the confusion matrix. **c**, The confusion matrix of the clustering by scDEC when $K=5$. The x-axis denotes where the cell is coming from while the y-axis denotes the inferred cluster. Mix CLP denotes CLP cells from both donors. **d**, The gap statistic shows two modes at $K=3$ and $K=5$. **e**, The visualization of the latent features learned by scDEC. The first ten dimensions correspond to the continuous latent variable $\tilde{\mathbf{z}}$ and the last three features correspond to the discrete latent variable $\tilde{\mathbf{c}}$. **f**, The clustering performance of scDEC when applying to a large mouse atlas dataset. **g**, The t-SNE visualization of around 10,000 PBMC cells coloured by the annotated labels from the 10x Genomics R&D team. **h**, The same t-SNE plot coloured by the normalized expression of the marker genes. **i**, The clustering performance of scDEC when applied to uni-modal single-cell data and multi-modal single-cell data (scRNA-seq and scATAC-seq measured in the same cell). The clustering performance of two methods were also demonstrated.

running time (Supplementary Fig. 15). We also noticed that the scDEC successfully identified most of the major reference cell types for each tissue (Supplementary Fig. 16).

scDEC enables integrative analysis of multi-modal single-cell data. It is natural to extend scDEC in multi-modal single-cell data analysis where multiple types of molecule within the same cell are measured simultaneously. Here, we apply scDEC to a dataset from 10x Genomics, which contains around 10,000 peripheral blood mononuclear cells (PBMC) with both measurements of scRNA-seq and scATAC-seq for each cell. Note that the granulocytes were removed by cell sorting of this dataset. After data preprocessing to scRNA-seq and scATAC-seq data, respectively, the two types of data are concatenated and fed into the scDEC model (see Methods). As the PBMC dataset has no FACS cell-type labels, we used the cell-type labels that were annotated by the 10x Genomics R&D team as surrogates. Most annotated cell types can be well distinguished by scDEC through the t-SNE visualization of the latent features (Fig. 4g). The visualization of different subpopulations of monocytes, T cells and B cells also demonstrates a clearer separation than using scRNA-seq or scATAC-seq only (Supplementary Fig. 17). The differentiable expression profiles of the several marker genes for PBMC cell types are illustrated in Fig. 4h. To name a few, *MS4A1* is a well-known marker gene for B cells[35], which is highly expressed in a cluster identified by scDEC. *FCER1A*, a marker gene for dendritic cells (DC)[36], is observed to be highly expressed in a tiny cluster identified by scDEC. Given surrogate cell labels, we evaluate the clustering performance of scDEC when applied to one type of data (scRNA-seq or scATAC-seq) and both types of single-cell data. scDEC achieves a substantially better clustering performance using both types of single-cell data than using scRNA-seq or scATAC-seq alone (Fig. 4i). Finally, we also compared scDEC to two recent methods on multi-modal single-cell data analysis. scDEC achieves an NMI of 0.779, ARI of 0.718 and homogeneity of 0.752, which outperforms MOFA+[37] and is comparable to scAI[38]. To sum up, scDEC can be easily extended to integrative analysis of multi-modal single-cell data analysis.

Discussion

In this study, we proposed scDEC for accurately characterizing cell subpopulations in scATAC-seq data using a deep generative model. Unlike previous studies that take dimension reduction and clustering as two independent tasks. scDEC intrinsically integrates the low-dimensional representation learning and unsupervised clustering together by carefully designing a GAN-based symmetrical architecture. scDEC can serve as a powerful tool for scATAC-seq data analysis, including visualization, clustering and trajectory analyses. In a series of experiments, scDEC achieves competitive or superior performance compared with other baseline methods. In downstream applications, we focused on the generation power of scDEC, which can facilitate the intermediate cell-state inference. The latent features learned by scDEC reveal both biological cell types and within-cell-type variations, which shed light on the biological mechanism. Our examples also showed that scDEC can handle very large datasets and is applicable to multi-modal single-cell data analysis.

We also provide several directions for improving scDEC. First, when applying scDEC to joint analysis of scRNA-seq and scATAC-seq data, it might be helpful to further enhance the clustering performance if scDEC model incorporates the relationship between genes and regulatory elements (REs). Second, the method of utilizing the generation power of scDEC can be further explored, especially in a complicated tree-based trajectory of cell differentiation or time-course single-cell profiles of cell development. Third, we note that there are already several tools or pipelines for single-cell batch-effect correction, such as Seurat-v3[39] and Harmony[40]. It is interesting to explore how to integrate such a procedure for data integrative analysis into scDEC models.

With scDEC, researchers could perform a scATAC-seq analysis or single-cell joint ATAC/RNA-seq analysis of the cell types or tissues with interests. Then one can simultaneously cluster single cells and uncover the biological findings underlying the learned latent features. We hope scDEC could help unveil the single-cell regulatory mechanism and contribute to understanding heterogeneous cell populations.

Methods

Data preprocessing. All the scATAC-seq datasets were uniformly preprocessed before being fed into the scDEC model. To reduce the level of noise, we only kept peaks with at least one read count in more than 3% of the cells. Next, similar to Cusanovich et al.[9], we applied a TF-IDF transformation to the raw scATAC-seq count matrix, which is a widely used technology in information retrieval and text mining[41,42]. We calculated the 'term frequency' by normalizing the raw reads count matrix for each cell through dividing the total reads count within that cell. The 'inverse document frequency' was calculated as the inverse frequency of each region to be accessible across all cells. The inverse document frequency was log-transformed and multiplied by the term frequency. The TF-IDF transformation helps increase proportionally to the number of times a peak appears in the cell, which gives a higher importance weight to the peaks with less frequency. Finally, a PCA[43] was applied to reduce the dimension of the scATAC to 20, which is implemented using the 'Scikit-learn' package[44]. scDEC shows robustness to the dimension of PCA (Supplementary Fig. 8). A summary of all scATAC-seq datasets used in this study is provided in Supplementary Table 4.

Visualization. We use t-SNE[18] as the default algorithm for visualizing the latent features of scATAC-seq data learned by different methods by setting the visualization dimension to 2. The t-SNE was implemented with the 'Scikit-learn' package[44]. The UMAP[19] was also implemented as an additional visualization tool for latent features.

Adversarial training in scDEC model. The scDEC model comprises a pair of GAN models. For forward GAN mapping, G network aims at conditionally generating samples $\{\widetilde{\mathbf{x}}_i\}_{i=1}^N$ that have a similar distribution to the observation data $\{\mathbf{x}_i\}_{i=1}^N$ while the discriminator $\mathrm{D_x}$ tries to discern observation data (positive) from generated samples (negative). The backward mapping function H and the discriminator $\mathrm{D_z}$ aim to transform the data from the data space to the latent space. Discriminators can be considered as binary classifiers where an input data point will be asserted to be positive (1) or negative (0). We use WGAN-GP[45] as the architecture for the GAN implementation, where the gradient penalty of discriminators will be considered as an additional loss term. We define the objective loss functions of the above four neural networks (G, H, $\mathrm{D_x}$ and $\mathrm{D_z}$) in the training process as the following

$$\begin{cases} \mathcal{L}_{\mathrm{GAN}}\left(\mathrm{G}\right) = -\underset{\mathbf{z}\sim p(\mathbf{z}),\mathbf{c}\sim \mathrm{Cat}(K,\mathbf{w})}{\mathbb{E}}\left[\mathrm{D}_{\mathbf{x}}(\mathrm{G}(\mathbf{z},\mathbf{c}))\right] \\ \mathcal{L}_{\mathrm{GAN}}\left(\mathrm{D}_{\mathbf{x}}\right) = -\underset{\mathbf{x}\sim p(\mathbf{x})}{\mathbb{E}}\left[\mathrm{D}_{\mathbf{x}}\left(\mathbf{x}\right)\right] + \underset{\mathbf{z}\sim p(\mathbf{z}),\mathbf{c}\sim \mathrm{Cat}(K,\mathbf{w})}{\mathbb{E}}\left[\mathrm{D}_{\mathbf{x}}\left(\mathrm{G}\left(\mathbf{z},\mathbf{c}\right)\right)\right] \\ \qquad +\lambda \underset{\widetilde{\mathbf{x}}\sim \hat{p}(\widetilde{\mathbf{x}})}{\mathbb{E}}\left[\left(\left|\left|\nabla_{\widetilde{\mathbf{x}}}\mathrm{D}_{\mathbf{x}}\left(\widetilde{\mathbf{x}}\right)\right|\right|_2 - 1\right)^2\right] \\ \mathcal{L}_{\mathrm{GAN}}\left(\mathrm{H}\right) = -\underset{\mathbf{x}\sim p(\mathbf{x})}{\mathbb{E}}\left[\mathrm{D}_{\mathbf{z}}\left(\mathrm{H}(\mathbf{x})\right)\right] \\ \mathcal{L}_{\mathrm{GAN}}\left(\mathrm{D}_{\mathbf{z}}\right) = -\underset{\mathbf{z}\sim p(\mathbf{z})}{\mathbb{E}}\left[\mathrm{D}_{\mathbf{z}}\left(\mathbf{z}\right)\right] + \underset{\mathbf{x}\sim p(\mathbf{x})}{\mathbb{E}}\left[\mathrm{D}_{\mathbf{z}}\left(\mathrm{H}\left(\mathbf{x}\right)\right)\right] \\ \qquad +\lambda \underset{\overline{\mathbf{z}}\sim \overline{p}(\overline{\mathbf{z}})}{\mathbb{E}}\left[\left(\left|\left|\nabla_{\overline{\mathbf{z}}}\mathrm{D}_{\mathbf{z}}\left(\overline{\mathbf{z}}\right)\right|\right|_2 - 1\right)^2\right] \end{cases}$$

where $p(\mathbf{z})$ and $\mathrm{Cat}(K, \mathbf{w})$ denote the probability distribution of continuous and discrete variables in the latent space, respectively. In practice, sampling $\mathbf{x}$ from $p(\mathbf{x})$ can be regarded as a procedure of randomly sampling from independent and identically distributed (i.i.d.) observation data with replacement. $\hat{p}\left(\hat{\mathbf{x}}\right)$ and $\overline{p}\left(\overline{\mathbf{z}}\right)$ denote uniformly sampling from the straight line between the points sampled from true data and generated data. Minimizing the loss of a generator (for example, $\mathcal{L}_{\mathrm{GAN}}\left(\mathrm{G}\right)$) and the corresponding discriminator (for example, $\mathcal{L}_{\mathrm{GAN}}\left(\mathrm{D}_{\mathbf{x}}\right)$) are somehow contradictory as the two networks (G and $\mathrm{D_x}$) compete with each other during the training process. λ is a penalty coefficient, which is set to 10 in all experiments.

Round-trip loss. During the training, we also aim to minimize the round-trip loss, which is defined as $\rho((\mathbf{z},\mathbf{c}), \mathrm{H}(\mathrm{G}(\mathbf{z},\mathbf{c})))$ and $\rho(\mathbf{x},\mathrm{G}(\mathrm{H}(\mathbf{x})))$ where $\mathbf{z}$ and $\mathbf{c}$ are sampled from the distribution of the continuous latent variable $p(\mathbf{z})$ and the category distribution $\mathrm{Cat}(K, \mathbf{w})$. The principle is to minimize the distance when a data point goes through a round-trip transformation between two data domains. In practice, we used l_2 loss as the continuous part of round-trip loss and cross-entropy loss as the discrete part in round-trip loss. We further denoted the roundtrip loss as

$$\mathcal{L}_{\mathrm{RT}}\left(\mathrm{G},\mathrm{H}\right) = \alpha\left\|\mathbf{x}-\mathrm{G}\left(\mathrm{H}\left(\mathbf{x}\right)\right)\right\|_2^2 + \alpha\left\|\mathbf{z}-\mathrm{H}_{\mathbf{z}}\left(\mathrm{G}\left(\mathbf{z},\mathbf{c}\right)\right)\right\|_2^2$$
$$+\beta\mathrm{CE}\left(\mathbf{c},\mathrm{H}_{\mathbf{c}}\left(\mathrm{G}\left(\mathbf{z},\mathbf{c}\right)\right)\right)$$

where α and β are two constant coefficients, which are both set to 10. $\mathrm{H_z}(\cdot)$ and $\mathrm{H_c}(\cdot)$ denote the continuous and discrete outputs from $\mathrm{H}(\cdot)$, respectively and $\mathrm{CE}(\cdot)$ represents the cross-entropy loss function. The idea of round-trip loss, which exploits transitivity for regularizing structured data, has also been used in previous work[16,46].

Full training loss. Combining the adversarial training loss and round-trip loss together, we can get the full training loss for generator networks and discriminator networks as $\mathcal{L}\left(\mathrm{G},\mathrm{H}\right) = \mathcal{L}_{\mathrm{GAN}}\left(\mathrm{G}\right) + \mathcal{L}_{\mathrm{GAN}}\left(\mathrm{H}\right) + \mathcal{L}_{\mathrm{RT}}\left(\mathrm{G},\mathrm{H}\right)$ and $\mathcal{L}\left(\mathrm{D}_{\mathbf{x}},\mathrm{D}_{\mathbf{z}}\right) = \mathcal{L}_{\mathrm{GAN}}\left(\mathrm{D}_{\mathbf{x}}\right) + \mathcal{L}_{\mathrm{GAN}}\left(\mathrm{D}_{\mathbf{z}}\right)$, respectively. To achieve joint training of the two GAN models, we iteratively updated the parameters in the two generative models (G and H) and the two discriminative models ($\mathrm{D_x}$ and $\mathrm{D_z}$), respectively. Thus, the overall iterative optimization problem can be represented as

$$\mathrm{G}^*,\mathrm{D}_{\mathbf{x}}^*,\mathrm{H}^*,\mathrm{D}_{\mathbf{z}}^* = \begin{cases} \arg\min\limits_{\mathrm{G,H}}\mathcal{L}\left(\mathrm{G},\mathrm{H}\right) \\ \arg\min\limits_{\mathrm{D_x,D_z}}\mathcal{L}\left(\mathrm{D}_{\mathbf{x}},\mathrm{D}_{\mathbf{z}}\right) \end{cases}$$

An Adam optimizer[47] with a learning rate of 2 × 10^{-4} was used for updating the weights in the neural networks. The training process is illustrated in detail in Supplementary Table 5.

Data generation in scDEC. We generate the state of intermediate cells by interpolating the latent indicator **c** of two 'neighbouring' cell types. Assume there are two cell types that correspond to the latent indicator $\mathbf{c}_1$ and $\mathbf{c}_2$, respectively. The generated data can be represented as $\mathrm{G}(\mathbf{z}, \tilde{\mathbf{c}})$ where $\tilde{\mathbf{c}} = \alpha\mathbf{c}_1 + (1 - \alpha)\mathbf{c}_2$. Note that the α is the generation coefficient from 0 to 1 and **z** is still sampled from a standard Gaussian distribution. The interpolation of latent features has already been used for exploring and visualizing the transition between two types of image[48].

Network architecture in scDEC. All the networks in scDEC are made of fully connected layers. The G network contains 10 fully-connected layers and each hidden layer has 512 nodes, while the H network contains 10 fully connected layers and each hidden layer has 256 nodes. $\mathrm{D}_\mathbf{x}$ and $\mathrm{D}_\mathbf{z}$ both contain 2 fully-connected layers and 256 nodes in the hidden layer. Batch normalization[49] was used in discriminator networks.

Updating the category distribution. The probability **w** in the Category distribution Cat(K, **w**) is adaptively updated every 100 batches of data based on the inferred cluster label from $\tilde{\mathbf{c}}$ of full training data (Supplementary Table 6).

Evaluation metrics for clustering. We compared different methods for clustering according to three metrics: normalized mutual information (NM)I[50], adjusted Rand index (ARI)[51] and homogeneity[52]. Assuming U and V are true and predicted label assignments given n data points, which have C_U and C_V clusters in total, respectively. NMI is then calculated by

$$\mathrm{NMI} = \frac{\sum_{p=1}^{C_U} \sum_{q=1}^{C_V} \left|U_p \cap V_q\right| \log \frac{n\left|U_p \cap V_q\right|}{\left|U_p\right| \times \left|V_q\right|}}{\max\left(-\sum_{p=1}^{C_U} \left|U_p\right| \log \frac{\left|U_p\right|}{n}, -\sum_{q=1}^{C_V} \left|V_q\right| \log \frac{\left|V_q\right|}{n}\right)}$$

The Rand index[53] is a measure of agreement between two cluster assignments while ARI corrects lacking a constant value when the cluster assignments are selected randomly. We define the following four quantities (1) n_1: number of pairs of objects in the same groups in both U and V, (2) n_2: number of pairs of objects in different groups in both U and V, (3) n_3: number of pairs of objects in the same group of U but different group in V, (4) n_4: number of pairs of objects in the same group of V but different group in U. Then ARI is calculated by

$$\mathrm{ARI} = \frac{\binom{n}{2}(n_1 + n_4) - [(n_1 + n_2)(n_1 + n_3) + (n_3 + n_4)(n_2 + n_4)]}{\binom{n}{2} - [(n_1 + n_2)(n_1 + n_3) + (n_3 + n_4)(n_2 + n_4)]}$$

Homogeneity is calculated by $\mathrm{Homo} = 1 - \frac{H(U|V)}{H(U)}$, where

$$\begin{cases} H(U|V) = -\sum_{p=1}^{C_U} \sum_{q=1}^{C_V} \frac{\left|U_p \cap V_q\right|}{n} \log \frac{\left|U_p \cap V_q\right|}{\sum_{q=1}^{C_V} \left|U_p \cap V_q\right|} \\ H(U) = -\sum_{p=1}^{C_U} \frac{\sum_{q=1}^{C_V} \left|U_p \cap V_q\right|}{C_U} \log \frac{\sum_{q=1}^{C_V} \left|U_p \cap V_q\right|}{C_U} \end{cases}$$

Estimating the number of clusters *K*. In order to apply scDEC to scATAC-seq where the number of cell types is unknown. We provide an algorithm for estimating the number of clusters K using a gap statistic[54]. We first compared the average within-cluster distance of the preprocessed scATAC-seq data and a reference dataset, which can be constructed with a random matrix with the same size using K-means algorithm. The average within-cluster distance on the reference dataset was calculated 1,000 times by Monto Carlo simulation and the average result was used. The optimal choice of K is given for which the gap between the single-cell data and the reference data is maximised. We note that this estimation of number of clusters K closely matches the truth clusters numbers with the scATAC-seq used in this study (Supplementary Fig. 18).

Identification of cluster-specific motifs and trajectory inference. The cluster-specific motifs are identified by Mann–Whitney U test[55] with the alternative hypothesis that the chromVAR scores[24] of cells in one cluster or multiple clusters have a positive shift compared with chromVAR scores of the rest of the cells. Then the motifs are ranked according to the p-values and the top-ranked motifs illustrated.

We used Slingshot[32] software with default parameters for trajectory inference. Given the latent features and the cell cluster labels inferred by scDEC, Slingshot is able to annotate smooth curves, which represent the estimated cell lineages.

Baseline methods. We compared scDEC to multiple baseline methods in this study, including scABC[7], SCALE[15], cisTopic[8], Scasat[10], Cusanovich2018[4,9] and SnapATAC[11]. SCALE was implemented from its original source code repository (https://github.com/jsxlei/SCALE). Other methods were implemented directly from a benchmark study[6]. For the methods (cisTopic, Scasat, Cusanovich2018 and SnapATAC) that only learn a low-dimension embedding of the scATAC-seq data, we used Louvain clustering[20], which was recommended by the benchmark study[6], as the default method for clustering the low-dimension embedding. Suggested by SCALE, we set the embedding dimension to the same number across different methods within a comparison experiment.

MOFA+[37] and scAI[38] are two recent works on multi-modal single-cell data analysis using matrix factorization frameworks. For MOFA+, we directly used the pretrained model on the same PBMC dataset, which can be downloaded from https://biofam.github.io/MOFA2/. scAI was implemented from its source code (https://github.com/sqjin/scAI) and the number of factors set to 20, which is the same as the dimension of latent features for scDEC. We applied K-means to the latent factors of MOFA+ and scAI in the clustering experiments. Note that the number of clusters K is set to 14, which is the number of cell types of the annotated label from the 10x Genomics R&D team.

Data preprocessing. Similar to SCALE, we filtered the scATAC-seq peaks by only keeping peaks that contain at least one read count in more than 3% of all cells. Uniform preprocessing demonstrated the robustness of the method across different scATAC-seq datasets. In the experiment of multi-modal single-cell analysis, we applied a uniform preprocessing strategy to scRNA-seq and scATAC-seq. We first filtered the genes or peaks that have zero read count across all cells. Then the read count matrix of scRNA-seq or scATAC-seq will be normalized in which the read count of each gene (peak) was divided by the total count in each cell and multiplied by a scale factor (10,000 by default). Next, a log-transformation was applied with a pseudocount of 1. At last, a PCA transformation was applied to scRNA-seq and scATAC-seq, respectively. The top 25 components of each type of data were kept and then concatenated together (50 in total) before being fed to scDEC.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The InSilico dataset was collected from the GEO database with accession number GSE65360. The mouse Forebrain dataset was downloaded from the GEO database with accession number GSE100033. The Splenocyte dataset can be accessed at ArrayExpress database with accession number E-MTAB-6714. The All blood dataset can be accessed at the GEO database with accession number GSE96772. The mouse atlas data are available at http://atlas.gs.washington.edu/mouse-atac. The human PBMCs dataset used in multi-modal single cell analysis was downloaded from 10x Genomics (https://support.10xgenomics.com/single-cell-multiome-atac-gex) with entry 'pbmc_granulocyte_sorted_10k'. The preprocessed scATAC-seq data used as input for scDEC model in this study can be downloaded from https://doi.org/10.5281/zenodo.3977858[56].

Code availability

scDEC is open-source software based on the TensorFlow library[57], which is available on Github (https://github.com/kimmo1019/scDEC) and Zenodo (https://doi.org/10.5281/zenodo.4560834)[58]. A CodeOcean capsule with several example datasets is available at https://codeocean.com/capsule/0746056/tree/v1[59]. The pretrained models on both benchmark single-cell datasets and 10x Genomics PBMCs multi-modal single-cell dataset were provided.

Received: 14 August 2020; Accepted: 14 March 2021;
Published online: 10 May 2021

References

1. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* **20**, 207–220 (2019).
2. Corces, M. R. et al. The chromatin accessibility landscape of primary human cancers. *Science* **362**, eaav1898 (2018).
3. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat. Rev. Genet.* **20**, 257–272 (2019).
4. Cusanovich, D. A. et al. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
5. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
6. Chen, H. et al. Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol.* **20**, 241 (2019).
7. Zamanighomi, M. et al. Unsupervised clustering and epigenetic classification of single cells. *Nat. Commun.* **9**, 2410 (2018).
8. González-Blas, C. B. et al. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods* **16**, 397–400 (2019).
9. Cusanovich, D. A. et al. A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell* **174**, 1309–1324.e1318 (2018).

10. Baker, S. M., Rogerson, C., Hayes, A., Sharrocks, A. D. & Rattray, M. Classifying cells with Scasat, a single-cell ATAC-seq analysis tool. *Nucleic Acids Res.* **47**, e10 (2019).
11. Fang, R. et al. Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat. Commun.* **12**, 1337 (2021).
12. Goodfellow, I. et al. Generative adversarial nets. In *Proceedings of Advances in Neural Information Processing Systems* (*NeurIPS*) 2672–2680 (NIPS, 2014).
13. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. In *Proceedings of International Conference on Learning Representations* (ICLR, 2014).
14. Liu, Q., Lv, H. & Jiang, R. hicGAN infers super resolution Hi-C data with generative adversarial networks. *Bioinformatics* **35**, i99–i107 (2019).
15. Xiong, L. et al. SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nat. Commun.* **10**, 4576 (2019).
16. Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision* 2223–2232 (ICCV, 2017).
17. Liu, Q., Xu, J., Jiang, R. & Wong, W. H. Density estimation using deep generative neural networks. *Proc. Natl Acad. Sci. USA* **118**, e2101344118 (2021).
18. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
19. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection. *J. Open Source Software* **3**, 861 (2018).
20. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).
21. Preissl, S. et al. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat. Neurosci.* **21**, 432–439 (2018).
22. Chen, X., Miragaia, R. J., Natarajan, K. N. & Teichmann, S. A. A rapid and robust method for single cell chromatin accessibility profiling. *Nat. Commun.* **9**, 5345 (2018).
23. Buenrostro, J. D. et al. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell* **173**, 1535–1548 (2018).
24. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
25. Mathelier, A. et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **44**, D110–115 (2016).
26. Shaltouki, A., Peng, J., Liu, Q., Rao, M. S. & Zeng, X. Efficient generation of astrocytes from human pluripotent stem cells in defined conditions. *Stem Cells* **31**, 941–952 (2013).
27. Bayam, E. et al. Genome-wide target analysis of NEUROD2 provides new insights into regulation of cortical projection neuron migration and differentiation. *BMC Genomics* **16**, 681 (2015).
28. Owa, T. et al. Meis1 coordinates cerebellar granule cell development by regulating Pax6 transcription, BMP signaling and Atoh1 degradation. *J. Neurosci.* **38**, 1277–1294 (2018).
29. Hallonet, M., Hollemann, T., Pieler, T. & Gruss, P. *Vax1*, a novel homeobox-containing gene, directs development of the basal forebrain and visual system. *Genes Dev.* **13**, 3106–3114 (1999).
30. Cesari, F. et al. Mice deficient for the Ets transcription factor Elk-1 show normal immune responses and mildly impaired neuronal gene activation. *Mol. Cell. Biol.* **24**, 294–305 (2004).
31. Stolt, C. C. et al. The Sox9 transcription factor determines glial fate choice in the developing spinal cord. *Genes Dev.* **17**, 1677–1689 (2003).
32. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
33. Iwasaki, H. & Akashi, K. Myeloid lineage commitment from the hematopoietic stem cell. *Immunity* **26**, 726–740 (2007).
34. Gilmour, J. et al. A crucial role for the ubiquitously expressed transcription factor Sp1 at early stages of hematopoietic specification. *Development* **141**, 2391–2401 (2014).
35. Anderson, K. C. et al. Expression of human B cell-associated antigens on leukemias and lymphomas: a model of human B cell differentiation. *Blood* **63**, 1424–1433 (1984).
36. Villani, A.-C. et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, eaah4573 (2017).
37. Argelaguet, R. et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 111 (2020).
38. Jin, S., Zhang, L. & Nie, Q. scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. *Genome Biol.* **21**, 25 (2020).
39. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e1821 (2019).
40. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
41. Teller, V. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. *Comput. Linguist.* **26**, 638–641 (2000).
42. Chowdhury, G. G. *Introduction to Modern Information Retrieval* (Facet, 2010).
43. Halko, N., Martinsson, P.-G. & Tropp, J. A. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.* **53**, 217–288 (2011).
44. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
45. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. & Courville, A. C. Improved training of Wasserstein GANs. In *Proceedings of Advances in Neural Information Processing Systems* 5767–5777 (NIPS, 2017).
46. Yi, Z., Zhang, H., Tan, P. & Gong, M. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision* 2849–2857 (ICCV, 2017).
47. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. In *Proceedings of International Conference on Learning Representations* (ICLR, 2014).
48. Mukherjee, S., Asnani, H., Lin, E. & Kannan, S. In *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 33, 4610–4617 (AAAI, 2019).
49. Ioffe, S. & Szegedy, C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning* 448–456 (ICML, 2015).
50. Strehl, A. & Ghosh, J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* **3**, 583–617 (2002).
51. Hubert, L. & Arabie, P. Comparing partitions. *J. Classification* **2**, 193–218 (1985).
52. Rosenberg, A. & Hirschberg, J. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* 410–420 (EMNLP-CoNLL, 2007).
53. Rand, W. M. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66**, 846–850 (1971).
54. Tibshirani, R., Walther, G. & Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. B* **63**, 411–423 (2001).
55. Mann, H. B. & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **18**, 50–60 (1947).
56. Liu, Q. et al. scDEC: data for simultaneous deep generative modeling and clustering of single cell genomic data. *Zenodo* https://doi.org/10.5281/zenodo.3984189 (2020).
57. Abadi, M. et al. Tensorflow: a system for large-scale machine learning. In *Proceedings of 12th USENIX Symposium on Operating Systems Design and Implementation* 265–283 (OSDI, 2016).
58. Liu, Q. et al. scDEC: code for simultaneous deep generative modeling and clustering of single cell genomic data. *Zenodo* https://doi.org/10.5281/zenodo.4560834 (2021).
59. Liu, Q. et al. scDEC: simultaneous deep generative modeling and clustering of single cell genomic data. *CodeOcean* https://doi.org/10.24433/CO.3347162.v1 (2020).

Acknowledgements

This work was supported by NIH grants R01 HG010359 (W.H.W.) and P50 HG007735 (W.H.W.). This work was also supported by the National Key Research and Development Program of China no. 2018YFC0910404 (R.J.), the National Natural Science Foundation of China nos 61873141 (R.J.), 61721003 (R.J.) and 61573207 (R.J.).

Author contributions

W.H.W., R.J. and Q.L. conceived the study. Q.L. designed and implemented scDEC. Q.L., S.C. and W.H.W. performed the data analysis. Q.L. and W.H.W. interpreted the results. Q.L., R.J. and W.H.W. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-021-00333-y.

Correspondence and requests for materials should be addressed to R.J. or W.H.W.

Peer review information *Nature Machine Intelligence* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Wing Hung Wong

Last updated by author(s): Mar 26, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
□	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
□	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used.
Data analysis	We used open-sourced software slingshot (https://github.com/kstreet13/slingshot) for annotating the cell trajectory.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

InSilico dataset was collected from GEO database with accession number GSE65360. The mouse forebrain dataset was downloaded from GEO database with accession number GSE100033. Splenocyte dataset can be accessed at ArrayExpress database with accession number E-MTAB-6714. All blood dataset can be accessed at GEO database with accession number GSE96772. The mouse atlas data is available at http://atlas.gs.washington.edu/mouse-atac. The human peripheral blood mononuclear cells (PBMCs) dataset used in multi-modal single cell analysis was downloaded from 10x Genomic website (https://support.10xgenomics.com/single-cell-multiome-atac-gex) with entry “pbmc_granulocyte_sorted_10k”. The preprocessed scATAC-seq data used as input for scDEC model in this study can be downloaded from https://doi.org/10.5281/zenodo.3977858.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	As we only use public datasets in our study. We used the original sample size from each public dataset.
Data exclusions	No data exclusion
Replication	Not relevant as we did conduct any wet biological experiments in our study.
Randomization	Not relevant as we did conduct any wet biological experiments in our study.
Blinding	Not relevant as we only use public datasets in our study

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging