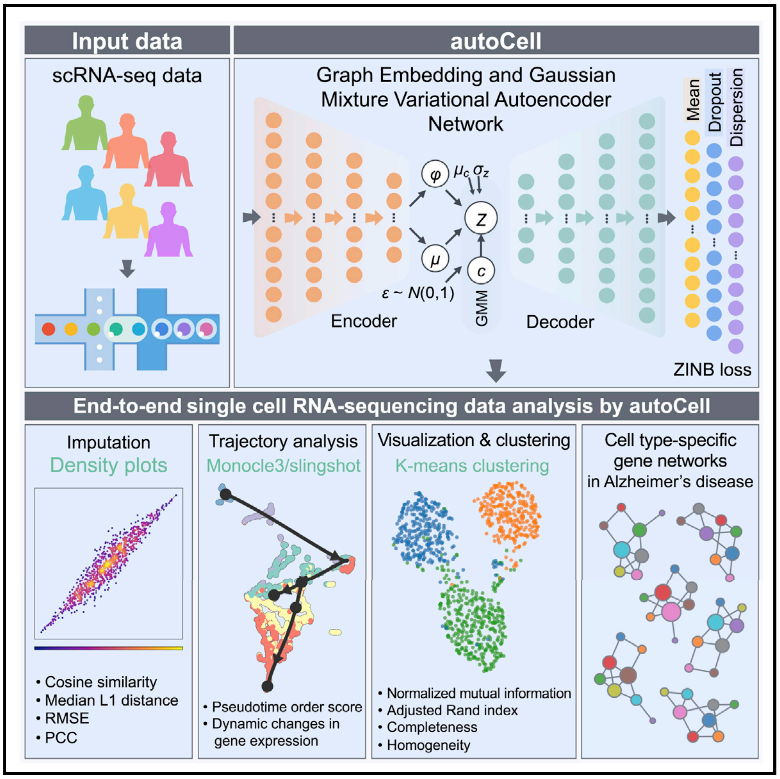
Cell Reports
Methods

Article

Graph embedding and Gaussian mixture variational autoencoder network for end-to-end analysis of single-cell RNA sequencing data

Graphical abstract



Highlights

- autoCell imputes heterogeneous and sparse sc/snRNA-seq data
- autoCell improves the performance of capturing cell developmental trajectories
- autoCell captures disease-relevant cellular pathobiology in latent space
- autoCell identifies cell-type-specific gene networks in Alzheimer's disease

Authors

Junlin Xu, Jielin Xu, Yajie Meng, ..., Xiangxiang Zeng, Ruth Nussinov, Feixiong Cheng

Correspondence

xzeng@hnu.edu.cn (X.Z.),
chengf@ccf.org (F.C.)

In brief

Xu et al. develop a graph-embedded Gaussian mixture variational autoencoder network algorithm (termed autoCell) for end-to-end analyses of single-cell/nuclei RNA-seq data, including visualization, clustering, imputation, and cell-type-specific gene network identification. autoCell offers a useful tool for large-scale single-cell genomic data analyses to accelerate translational biology and disease discoveries.

Xu et al., 2023, Cell Reports Methods *3*, 100382
January 23, 2023 © 2022 The Author(s).
https://doi.org/10.1016/j.crmeth.2022.100382

Cell Reports Methods

CellPress
OPEN ACCESS

Article

Graph embedding and Gaussian mixture variational autoencoder network for end-to-end analysis of single-cell RNA sequencing data

Junlin Xu,[1] Jielin Xu,[2] Yajie Meng,[1] Changcheng Lu,[1] Lijun Cai,[1] Xiangxiang Zeng,[1,*] Ruth Nussinov,[3,4]
and Feixiong Cheng[2,5,6,7,*]

[1]College of Computer Science and Electronic Engineering, Hunan University, Changsha, Hunan 410082, China
[2]Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44195, USA
[3]Computational Structural Biology Section, Basic Science Program, Frederick National Laboratory for Cancer Research, National Cancer Institute at Frederick, Frederick, MD 21702, USA
[4]Department of Human Molecular Genetics and Biochemistry, Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel
[5]Department of Molecular Medicine, Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, OH 44195, USA
[6]Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, Cleveland, OH 44106, USA
[7]Lead contact
*Correspondence: xzeng@hnu.edu.cn (X.Z.), chengf@ccf.org (F.C.)

https://doi.org/10.1016/j.crmeth.2022.100382

MOTIVATION Single-cell RNA sequencing (scRNA-seq) enables researchers to study gene expression at cellular resolution. However, noise caused by amplification and dropout may hamper precise data analyses. It is urgent to develop scalable denoising methods to deal with the increasingly large, but sparse, scRNA-seq data. Here, we present autoCell, a graph-embedded Gaussian mixture variational autoencoder network algorithm for scRNA-seq dropout imputation and feature extraction. Our autoCell provides a deep-learning toolbox for end-to-end analysis of large-scale single-cell/nucleus RNA-seq data, including visualization, clustering, imputation, and disease-specific gene network identification.

SUMMARY

Single-cell RNA sequencing (scRNA-seq) is a revolutionary technology to determine the precise gene expression of individual cells and identify cell heterogeneity and subpopulations. However, technical limitations of scRNA-seq lead to heterogeneous and sparse data. Here, we present autoCell, a deep-learning approach for scRNA-seq dropout imputation and feature extraction. autoCell is a variational autoencoding network that combines graph embedding and a probabilistic depth Gaussian mixture model to infer the distribution of high-dimensional, sparse scRNA-seq data. We validate autoCell on simulated datasets and biologically relevant scRNA-seq. We show that interpolation of autoCell improves the performance of existing tools in identifying cell developmental trajectories of human preimplantation embryos. We identify disease-associated astrocytes (DAAs) and reconstruct DAA-specific molecular networks and ligand-receptor interactions involved in cell-cell communications using Alzheimer's disease as a prototypical example. autoCell provides a toolbox for end-to-end analysis of scRNA-seq data, including visualization, clustering, imputation, and disease-specific gene network identification.

INTRODUCTION

Single-cell technology is a revolutionary breakthrough, allowing us to study the genome, transcriptome, and multi-omics systems of each cell, in each state, in tissues.[1,2] Combined with technologies such as fluorescent labeling and microdissection, it can also determine spatial attributes and cell-cell communication. These technologies have been widely used, leading to a revolution in basic and translational medicine.

Single-cell or single-nucleus RNA sequencing (sc/snRNA-seq) is important for identifying biological and disease-relevant cell types and subpopulations from heterogeneous cells.[3–6] Low-dimensional analysis of expression in different cell states can also be highly effective in reconstructing the cell developmental trajectory.[7–12] However, the amount of mRNA in a single cell is small, which necessitates a nearly million-fold amplification. Although the measurement technology has been greatly improved, technical factors still cause considerable noise in data generated in scRNA-seq experiments, including amplification deviation, library size difference, and extremely low capture rate. In particular, the extremely low RNA capture rate leads to undetectable, albeit expressed, genes, namely "dropout" events.[13] An essential difference is found between the "false" zero count caused by the "dropout" event and the true zero count. Given the sparse expression metrics, traditional analytic tools cannot achieve scientific rigor, and they lack high data reproducibility.[14]

Deep-learning algorithms have shown compelling performance in high-dimensional data processing, including sparse genomic data. DeepImute, an interpolation algorithm based on a deep neural network, is one such example. It uses the dropout layer to learn the patterns in the data to achieve accurate imputation.[15] Another denoising model is to denoise scRNA-seq datasets through a depth-counting autoencoder (DCA) network. A DCA takes the count distribution, overdispersion, and sparsity of the data into account using a negative binomial noise model with zero inflation.[16] scScope, a scalable deep-learning-based method, can accurately and quickly identify cell-type composition from millions of heterogeneous single-cell transcriptomic profiles.[9] Compared with DCA and scScope, scVI[17] uses a variational autoencoder (VAE) to reduce the dimensionality of scRNA-seq data. However, the standard VAE implemented by scVI[17] only uses a single isotropic multi-variate Gaussian distribution on the latent variable, which is not generally suitable for representing multi-category data such as scRNA-seq data containing multiple types of cells/nuclei. scVAE is another method based on VAEs, which uses Gaussian mixture models (GMMs) as prior distributions and introduces Poisson or negative binomial distributions to obtain latent representations of cells.[18] However, the model considers the mean and variance of the GMM as random variables, and it is approximated by a neural network with a parameter β, which makes model optimization challenging. As the number of measurable cells/nuclei and additional emerging sc/snRNA-seq data analysis challenges increase, the demand for faster and scalable estimation methods becomes a pressing need.

With the rapid development of graph neural networks, graph autoencoders can learn low-dimensional representations of graph topology and train node relationships with a global, entire graph view. Increasingly, exploiting a graph neural network framework in sc/snRNA-seq data analysis can be considered. Single-cell graph neural network (scGNN)[19] establishes cell-cell relationships through graphical neural networks and uses left-truncated mixed Gaussian models to model heterogeneous gene expression patterns. It also integrates three iterative multi-mode autoencoders. However, the iterative autoencoder framework requires more computing resources, which is more time consuming.

In this study, we present a deep-learning framework, namely autoCell, for dropout imputation and feature extraction from scRNA-seq data. The accuracy index indicates that autoCell outperformed six state-of-the-art published imputation methods in simulated datasets and biologically relevant sc/snRNA-seq datasets with varying degrees of human diseases. Therefore, autoCell is a scalable and accurate scRNA-seq data processing method that is superior to other scRNA-seq data analysis tools.

RESULTS

Overview of the autoCell framework

The overview of autoCell is shown in Figure 1. It is a VAE network that combines graph embedding and GMMs to model the distribution of high-dimensional, sparse scRNA-seq data. autoCell architecture can use biological representations of cells and genes to perform different scRNA-seq data analysis tasks. By integrating GMMs, autoCell can better estimate data distribution. We apply graph embedding to deal with sc/snRNA-seq data. Capturing the graphical information of the local data structure is a good complement to deep GMMs, making the network learn a global model with local structure constraints. Recent studies have showed that zero-inflated negative binomial (ZINB) distribution for modeling is an appropriate tool to solve the "dropout" event for scRNA-seq data. In reducing the impact of dropout events in highly sparse and over-dispersed count data, we introduce the ZINB distribution model, thereby denoising scRNA-seq data (see STAR Methods).

autoCell effectively imputes scRNA-seq data

We first applied autoCell to simulated scRNA-seq data to assess its imputation performance.[20] We simulated two datasets with three cell types, including 1,000 cells and 2,000 genes. For the simulation of the two datasets, 60% and 71% of the data values were set to the zero matrix to simulate dropout events in real data. We divided the entry of the simulated raw expression data into zero and non-zero space. Based on the density plot of the estimated and real values, the recovery values of DCA and autoCell are closer to the true expression values, and scGNN is at a medium level. MAGIC,[21] SAVER,[22] and SAUCIE[23] always tend to underestimate the original value. We also calculated the median L1 distance, root-mean-square error (RMSE) scores, and cosine similarity (see STAR Methods) score between the real expression value and restored expression value to measure the estimation accuracy. As shown in Figure 2, autoCell achieves overall better performance than the others. In particular, autoCell ranks second in the median L1 distance of gene expression restoration on the two simulated datasets and ranks second in the cosine similarity score on the simulated dataset with a synthetic dropout rate of 71%. In addition, the dropout event will increase the noise, muddling the identity of cell types, which can be restored through interpolation by value recovery algorithms. We also found that autoCell outperforms existing methods in the two simulation datasets (Figures S1A and S1B).

In evaluating the performance of autoCell in imputing missing values, we also selected two biologically relevant sc/snRNA-seq datasets[24,25] with well-annotated cell types as benchmarks. We simulated the dropout effects by randomly flipping 10% non-zero entries to the zero matrix. Similarly, three indicators between the original dataset and imputed values of these synthetic items are calculated as a measure of estimation accuracy. Compared with several state-of-the-art algorithms (Figure 2), autoCell achieves the best performance evaluated by the median

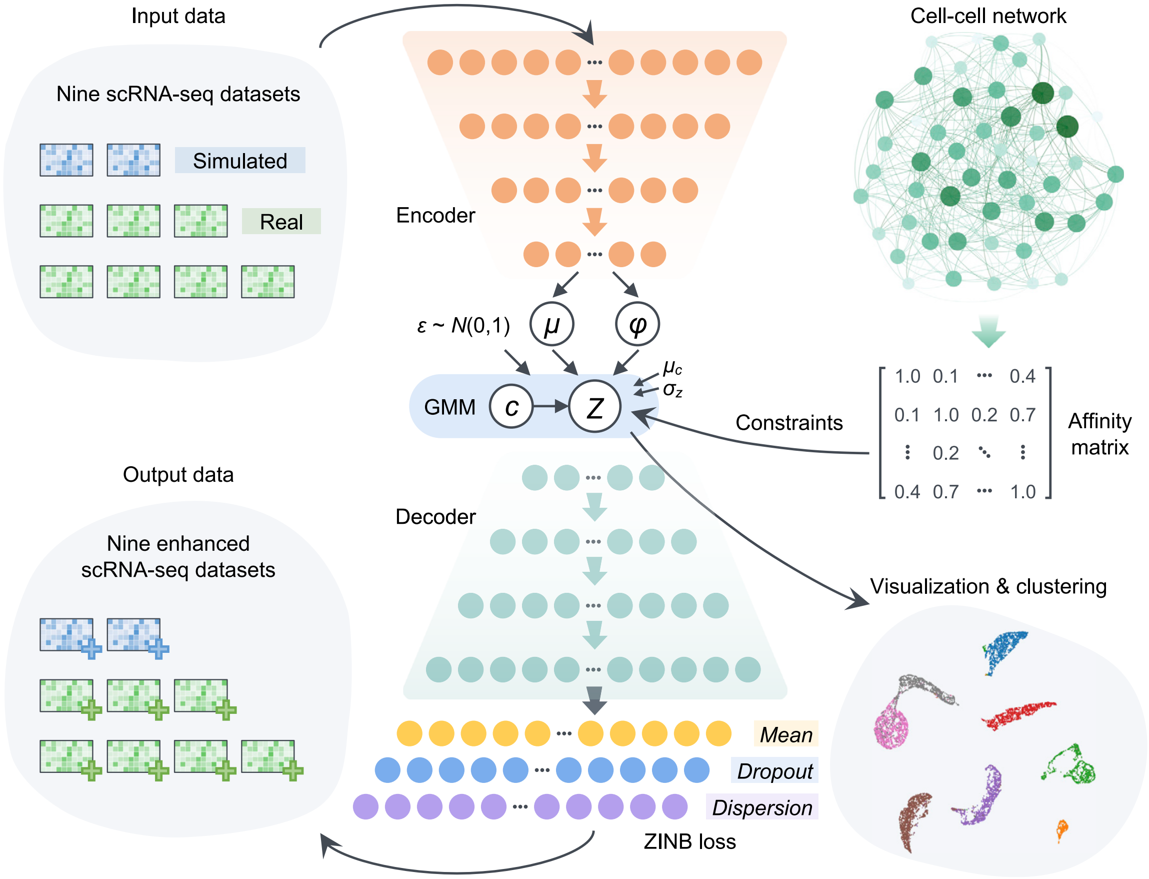


Figure 1. Overview of the autoCell framework

autoCell uses a Gaussian mixture model (GMM) and a deep neural network (DNN) to model the process of data generation. It uses zero-inflated negative binomial (ZINB) loss to process "dropout" events in scRNA-seq. The encoder and decoder are a two-layer neural network (128–128) with 10-dimensional latent variables (features) directly connected to the output. The cell-cell network is used to constrain the latent feature Z; thus, similar cells have similar latent features and cluster assignments. The yellow nodes depict the mean of the negative binomial distribution, which is the main output of the method representing denoised data, whereas the purple and blue nodes represent the other two parameters of the ZINB distribution, namely, dispersion and dropout.

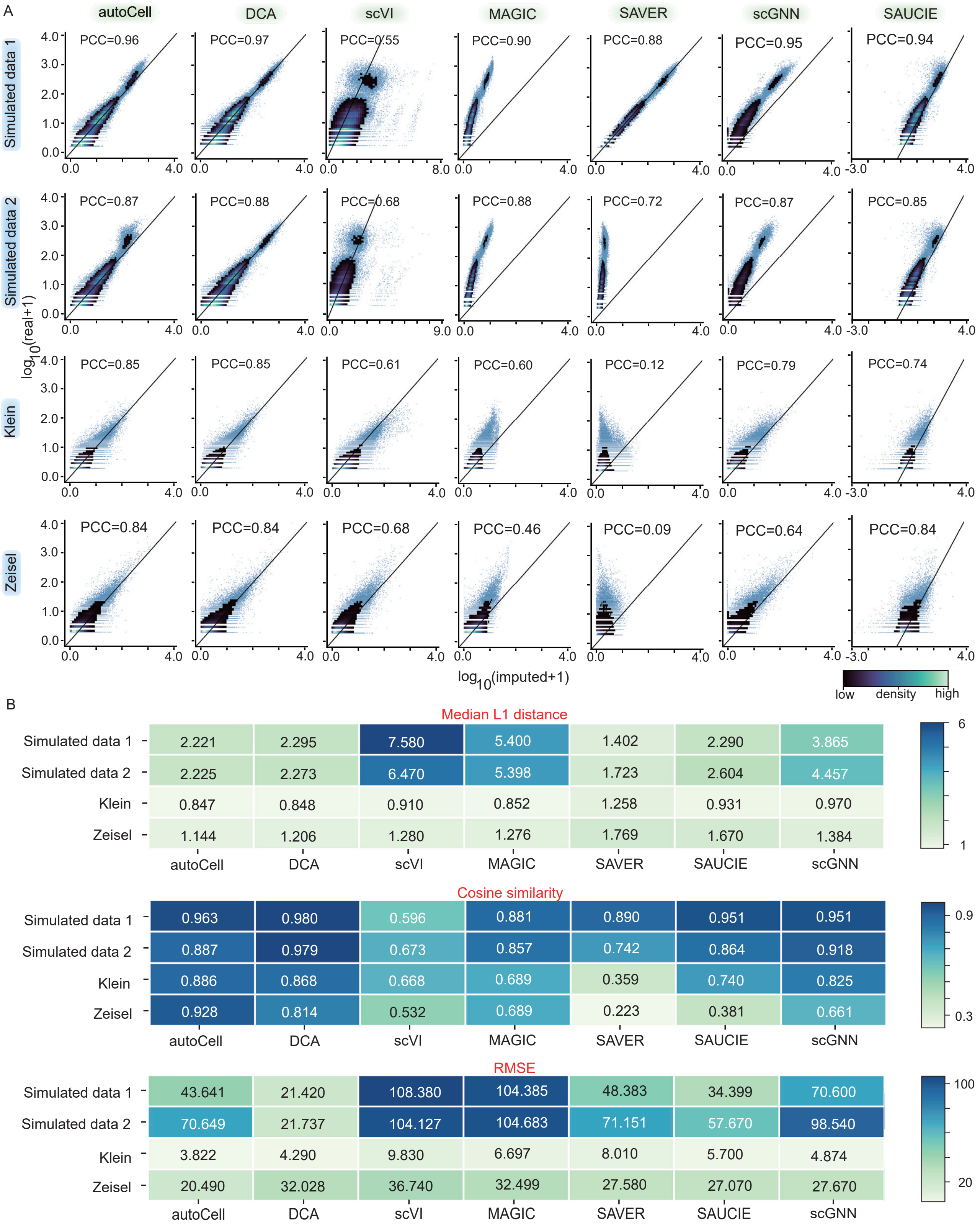
L1 distance, cosine similarity, and RMSE at the 10% synthetic dropout rate. Furthermore, autoCell imputation is closer to the true expression value based on the density plot of the estimated and true values (Figure 2). Collectively, autoCell outperforms state-of-the-art methods in sc/snRNA-seq data imputation analysis (Figure S2).

autoCell improves the performance of existing tools for capturing cell developmental trajectories

Apart from identifying cell types, scRNA-seq facilitates the organization of cells by time course or developmental stage (i.e., cell trajectory). The transition of cells from one functional state to another is a critical event in development. However, transitions are difficult to characterize since trapping and purifying cells in between stable endpoint states are challenging. Although some models currently exist to infer cell developmental trajectories based on scRNA-seq data, most inference methods do not address dropout events. We tested the accuracy of inferring the cell trajectory of scRNA-seq data after interpolation via autoCell. We used a benchmark dataset[26] with 1,529 single cells with five stages of well-annotated human preimplantation embryonic development from embryogenesis embryonic day 3 (E3) to E7. We reconstructed cell trajectories using monocle3[27] after various interpolation processes. The interpolation of autoCell produced the highest correspondence between the inferred pseudotime and the real-time cellular development (Figure S3). Pseudotime order score (POS; see STAR Methods) increased from 0.838 to 0.850. On the contrary, the POS obtained by other algorithms is lower than the original scRNA-seq dataset (Figure S3). Furthermore, we used another common trajectory analysis model, slingshot,[28,29] to test whether autoCell improves trajectory analysis. We found that cell development trajectory is well captured by interpolation from autoCell (Figure 3). Therefore, autoCell captures more accurate transcriptome dynamics and cell developmental trajectories across different developmental stages.

autoCell captures cellular pathobiology in latent space

We also assessed the extent to which the latent space inferred by autoCell reflects the biological variability among cells based on the previous stratification of cells into biologically important



Median L1 distance

	autoCell	DCA	scVI	MAGIC	SAVER	SAUCIE	scGNN
Simulated data 1	2.221	2.295	7.580	5.400	1.402	2.290	3.865
Simulated data 2	2.225	2.273	6.470	5.398	1.723	2.604	4.457
Klein	0.847	0.848	0.910	0.852	1.258	0.931	0.970
Zeisel	1.144	1.206	1.280	1.276	1.769	1.670	1.384

Cosine similarity

	autoCell	DCA	scVI	MAGIC	SAVER	SAUCIE	scGNN
Simulated data 1	0.963	0.980	0.596	0.881	0.890	0.951	0.951
Simulated data 2	0.887	0.979	0.673	0.857	0.742	0.864	0.918
Klein	0.886	0.868	0.668	0.689	0.359	0.740	0.825
Zeisel	0.928	0.814	0.532	0.689	0.223	0.381	0.661

RMSE

	autoCell	DCA	scVI	MAGIC	SAVER	SAUCIE	scGNN
Simulated data 1	43.641	21.420	108.380	104.385	48.383	34.399	70.600
Simulated data 2	70.649	21.737	104.127	104.683	71.151	57.670	98.540
Klein	3.822	4.290	9.830	6.697	8.010	5.700	4.874
Zeisel	20.490	32.028	36.740	32.499	27.580	27.070	27.670

(legend on next page)

subpopulations through unsupervised clustering followed by manual inspection and annotation. We applied autoCell to two simulated datasets and four biologically relevant scRNA-seq datasets. The zero ratio of these six datasets ranges from 60% to 90%. By default, autoCell extracted 10 features from the input data. For a fair comparison, we further applied common scRNA-seq data dimension reduction methods, including scVI, DESC,[30] scVAE, DCA, and SAUCIE, to reduce the input data to 10 dimensions. We used uniform manifold approximation and projection (UMAP) to visualize the features extracted from these tools and the original data. We found that the feature embeddings of autoCell were well separated among cell types with closer inner-group distances and larger between-group distances. However, the embedding and original data of DESC, scVAE, DCA, and SAUCIE overlapped among certain cell types. We found that autoCell is the best approach to identify all three cell types in the two simulated datasets (Figure S1C). For the Klein dataset,[25] scVI, scVAE, and autoCell showed better performance, and DCA caused the cell types d0 and d2 to be closely linked. However, SAUCIE and DESC only separated cells with cell type d0 and incorrectly divided cell type d7 into two cell types (Figure 4A). For the Zeisel dataset,[24] we found that autoCell, scVI, and scVAE still outperformed the other models, and autoCell and scVAE achieved a closer intra-group distance (Figure 4B).

We applied K-means clustering on autoCell-extracted latent features and assessed the clustering accuracy by comparing it with scVI, DESC, scVAE, DCA, and SAUCIE. We found that autoCell displays the best performance across all tested scRNA-seq datasets (Figure 4). In the Klein dataset,[25] the clustering output using autoCell (Figure 4C) was more consistent with the predefined unit-type annotation (normalized mutual information [NMI] = 0.882 and adjusted Rand index [ARI] = 0.907) than the second ranked model, scVI (NMI = 0.832, ARI = 0.784). In the Zeisel dataset,[24] the clustering performance of autoCell was considerably better than other existing tools. Then, we changed K-means to another four clustering algorithms (including spectral clustering, affinity propagation, birch, and agglomerative clustering). autoCell shows the best performance across all tested scRNA-seq datasets (Table S1).

In addition, we compared the visualization performance using principal-component analysis (PCA). As shown in Figure S4, in the Klein dataset, the feature embedding of autoCell, scVI, and scVAE was well separated among cell types. For the Zeisel dataset, autoCell and DESC showed the best performance, although they mixed astrocytes and brain endothelial cells, both of which were glial cells located in the CNS, but they can effectively separate most cell types. For the Romanov dataset, all models identify two glial clusters (astrocytes and brain endothelial cells) close to each other in the latent space. However, autoCell was the only model that effectively separated the microglia (Figure S4).

Collectively, autoCell presented elevated accuracy in capturing cellular pathobiology than existing state-of-the-art approaches for simulated and real-world biologically relevant scRNA-seq datasets (Figures S4 and S5).

Discovery of cell-type-specific molecular networks by autoCell

In testing whether the autoCell-inferred cell type can capture specific pathobiology of human diseases, we analyzed astrocytes, microglia, neurons, and oligodendrocyte progenitor cells (OPCs) using Alzheimer's disease (AD) as a prototypical example. In total, we re-analyzed 13,214 high-quality nuclei generated from the entorhinal cortex of AD brains and healthy controls. Using autoCell, we identified four microglia clusters, nine astrocyte clusters, and five OPC clusters (Figure 5A). Recent studies using human postmortem brain tissues identified disease-associated astrocyte (DAA) involved in crucial roles of AD pathogenesis and disease progression of AD. Using 11 experimentally validated DAA marker genes (four upregulated marker genes [*GFAP*, *CD44*, *HSPB1*, and *TNS*] and seven downregulated marker genes [*SLC1A2*, *SLC1A3*, *GLUL*, *NRXN1*, *CADM2*, *PTN*, and *GPC5*]),[31] we identified astrocyte subcluster 4 as a DAA by autoCell (Figure S6). Next, we built a DAA-specific molecular network using a state-of-the-art network-based algorithm, GPSnet,[32] under the human protein-protein interaction (PPI) network model. The DAA-specific module network included 50 PPIs connected by 44 proteins, such as APOE, MAPT, CD44, FOS, and STAT3 (Figure 5B; Table S2). *APOE* and microtubule-associated protein Tau (*MAPT*) were two of the most well-known risk genes for AD.[33,34] CD44 was an inflammation-associated protein. The inhibition of CD44 could be a potential strategy for AD treatment.[35] In a mouse model study, Stat3-deficient and Stat3-deletion astrocytes presented dropped levels of β-amyloid and pro-inflammatory cytokine activities.[36] Proteins from the DAA-specific molecular network were enriched by multiple AD-related pathways, such as cytokine signaling, spinal cord injury, and brain-derived neurotrophic factor signaling pathways (Figure 5B; Table S3). For example, several proteins in the DAA-specific network (STAT3, MAPT, HSPB8, HSPB1, JUNB, and LINGO1) were enriched in multiple cytokine signaling pathways, including interleukin-5 (IL-5), IL-2, IL-18, IL-3, and IL-4, consistent with the important role of neuroinflammation mediated by microglia in AD.[37,38] Therefore, using autoCell, we can identify disease-associated, cell-type-specific molecular network involved in key AD pathobiology.

We also identified significant ligand-receptor interactions involved in cell-cell communications in AD. We first inferred cell subpopulations using autoCell and the predicted ligand-receptor interaction using CellChat.[39] As shown in Figure 5C, we found strong ligand-receptor interactions among astrocytes, OPCs, and oligodendrocytes compared with the other three cell types

Figure 2. Performance comparison between autoCell and other state-of-the-art methods under 10% synthetic dropout rate

(A) Density plots of imputed versus original data masked. The x axis corresponds to the imputed values, and the y axis represents the true values of the masked data points. Each row is a different dataset, and each column is a different imputation method. Pearson correlation coefficient (PPC; higher is better).

(B) Comparison of the cosine similarity (higher is better), median L1 distance (lower is better), and root-mean-square error (RMSE) scores (lower is better) between autoCell and the other six imputation tools.

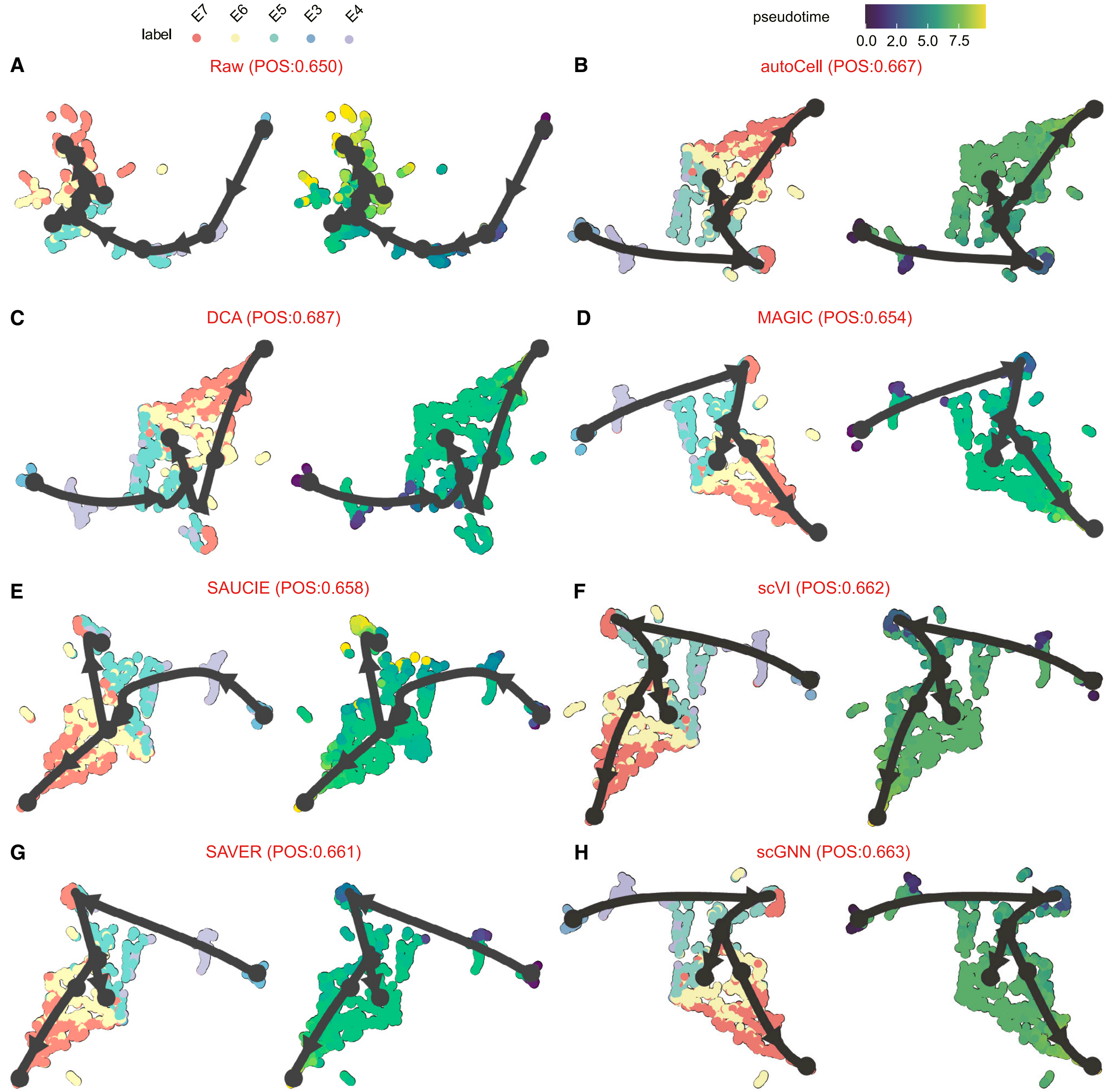


Figure 3. autoCell improves pseudotime analysis in the human preimplantation embryonic development dataset

(A and B) Results of slingshot estimated pseudotime using (A) raw data as input and (B) processed data from autoCell as input.
(C) Processed data from DCA as input.
(D) Processed data from MAGIC as input.
(E) Processed data from SAUCIE as input.
(F) Processed data from scVI as input.
(G) Processed data from SAVER as input.
(H) Processed data from scGNN as input.
POS, pseudotime order score (the higher the value the better). Pseudotime is a measure of how much progress an individual cell has made through a process such as cell differentiation.

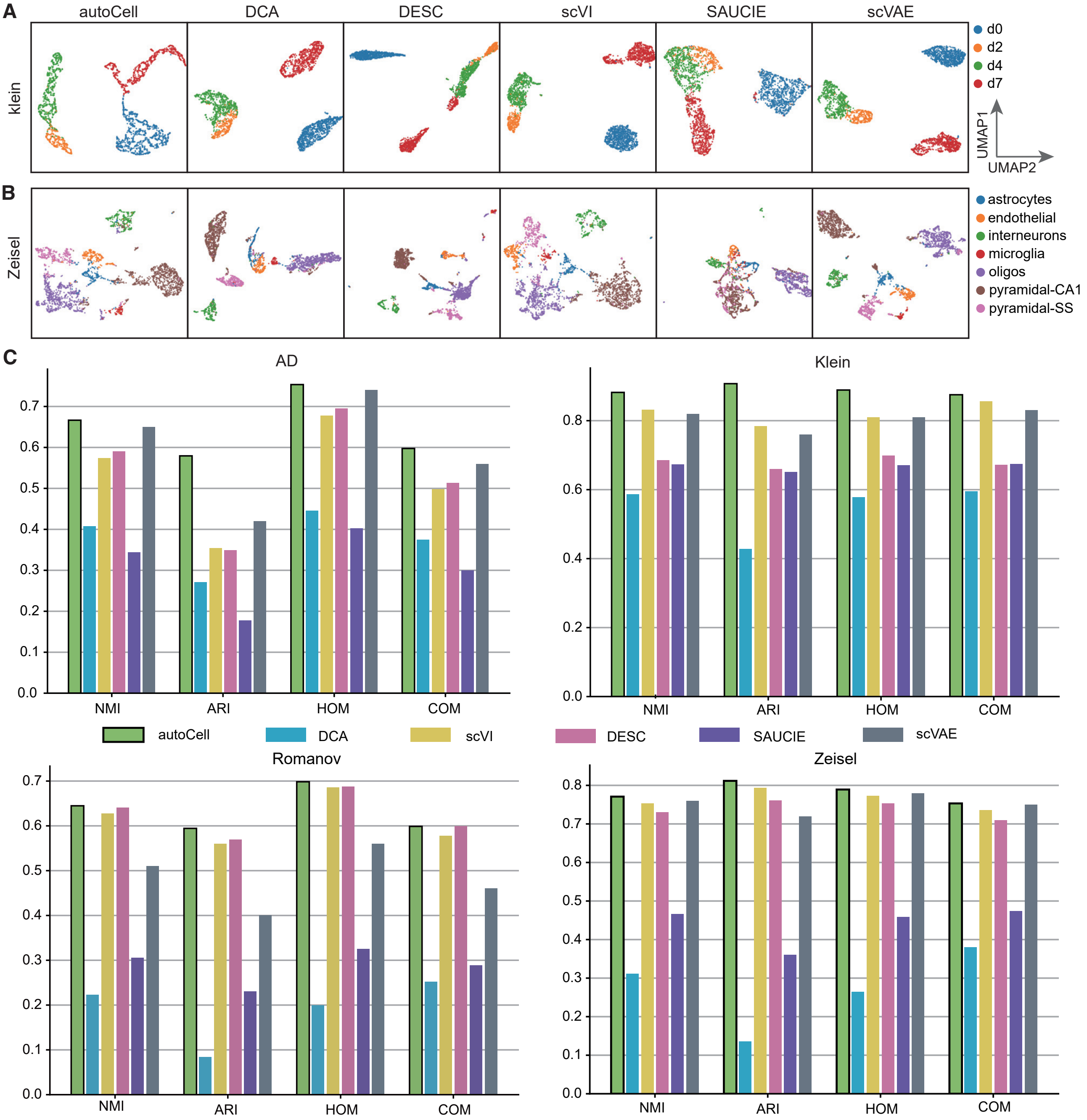


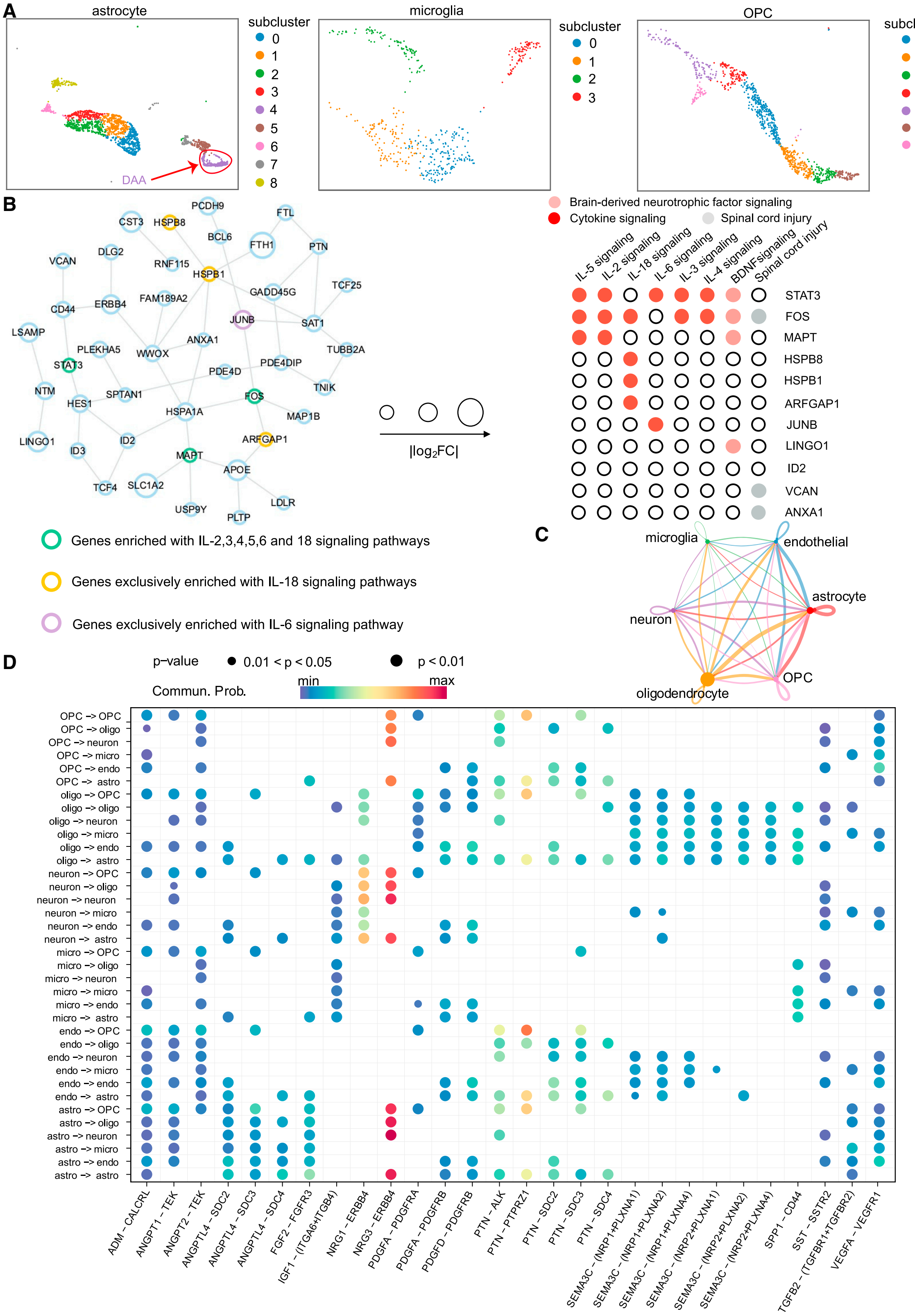
Figure 4. UMAP visualization of the extracted features using different approaches

(A and B) We evaluated autoCell with DCA, DESC, scVI, SAUCIE, and scVAE using two datasets: (A) Klein and (B) Zeisel datasets. For comparison, autoCell, DCA, DESC, scVI, SAUCIE, and scVAE all performed dimension reduction to 10 dimensions before applying UMAP.

(C) Comparison on the effect of clustering on four benchmark datasets. Clustering accuracy was evaluated by applying K-means clustering on the extracted features to obtain cluster assignments.

NMI, normalized mutual information (the higher the value the better); ARI, adjusted rand index; COM, completeness (the higher the value the better); HOM, homogeneity (the higher the value the better).

(neuron, microglia, and endothelial). Two ligand-receptor pairs (NRG3-ERBB4 and NRG1-ERBB4) revealed strong interactions across multiple cell-cell pairs (Figure 5D; Table S4). Multiple single-nucleotide polymorphisms in the NRG3 gene were found to be associated with the onset of AD.[40] In addition, the overexpression of ERBB4 in neurons was found to be associated with



(legend on next page)

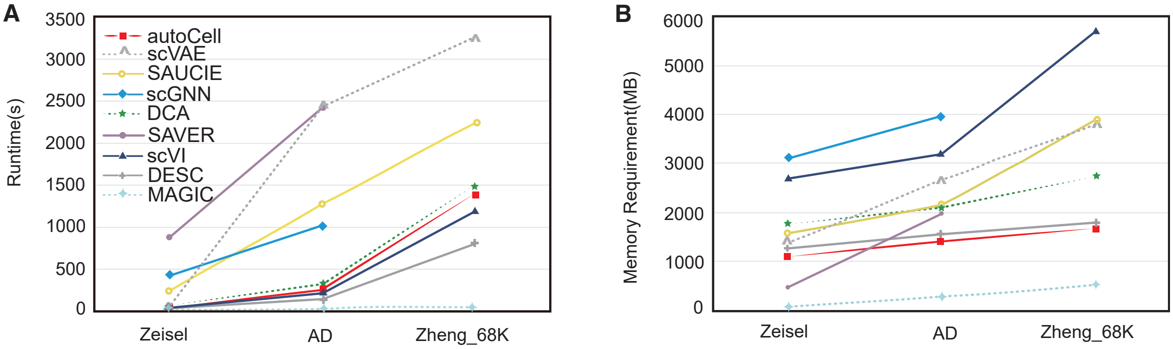


Figure 6. Comparison of running time and memory usage

autoCell scales linearly with the number of cells. The plot shows the running time and memory on three different size datasets: (1) Zeisel (3,005 cells); (2) AD dataset (13,214 cells/nuclei); and (3) Zheng-68k dataset (68,579 cells). Colors indicate different methods. The experimental environment is Intel Core i9-10900K CPU@3.70GHz, NVIDIA RTX 3080.

AD neuropathology.[41] A recent AD mouse model study found that immunoreactivity of NRG1 and ERBB4 was associated with plaques in the hippocampus region.[42] Using AD as a prototypical example, we demonstrated that the disease-associated cell subtype identified by autoCell could identify molecular targets and networks (i.e., ligand-receptor interactions) involved in AD pathogenesis and provide potential drug targets for AD or other human diseases if broadly applied.

autoCell is scalable to large datasets

The increasing number of cells is the main challenge in scRNA-seq analysis. In large projects such as the Human Cell Atlas,[43] the number of cells may be hundreds of thousands. In large datasets, identifying cell populations is challenging because many existing scRNA-seq clustering methods cannot scale up to handle them. Thus, we used the Zheng-68k[44] dataset and the Zheng-73k[44] dataset to study whether autoCell is suitable for large datasets. These cell types were used as references when benchmarking autoCell. autoCell worked well on these two large datasets and obtained good clustering performance (Figures S2 and S4). In addition, we performed time and memory comparisons on three datasets with different sizes: (1) Zeisel (3,005 cells), (2) AD dataset (13,214 cells), and (3) Zheng-68k dataset (68,579 cells). As shown in Figure 6, autoCell ranks fourth among all algorithms with regard to time. With regard to memory, autoCell ranks second after MAGIC. Thus, autoCell is scalable to large datasets and is comparable to computing time and memory compared with existing methods.

DISCUSSION

Single-cell technology is a revolutionary breakthrough that leads to the study of genomic, transcriptomic, and multi-omics systems of each cell in each tissue for the first time. With the development and wide application of technology, computational methods have been developed to solve problems posed by its generated data. Among these methods, dimensionality reduction (or low-dimensional representation) is the basis of scRNA-seq data visualization and downstream analysis. However, the technical shortcoming of single-cell sequencing and the transcription burst effect in single cells cause the data to be noisier than bulk RNA-seq data. Among the problems, dropout events usually occur, in which the false value of genes in some cells is zero or close to zero, limiting the performance of dimensionality reduction.

Here, we proposed a deep model autoCell for feature extraction and dropout imputation of scRNA-seq data. The key innovation of autoCell is the use of GMMs to estimate the latent feature distribution of the data. Compared with the previous application of the VAE in scRNA-seq data analysis, autoCell captures the graphical information of the local data structure by introducing graph embedding. This is an excellent supplement to the deep GMM, which allows network learning to follow the global model with local structure constraints. In reducing the impact of dropout events, we introduced the ZINB distribution, which can model highly sparse and over-dispersed count data, thereby denoising scRNA data. Through systematic comparison between simulated and real datasets, autoCell achieves better

Figure 5. Discovery of cell-type-specific molecular networks and significant ligand-receptor interactions in Alzheimer's disease (AD) using autoCell

(A) UMAP visualization of subclusters of astrocytes (including disease-associated astrocyte [DAA]), microglia, and OPCs, labeled with autoCell clusters.

(B) Reconstruction of the DAA-specific molecular subnetwork from the human protein-protein interactome. The DAA-specific module network included 50 protein-protein interactions (PPIs) connected by 44 proteins (e.g., APOE, MAPT, CD44, FOS, and STAT3). The proteins from the DAA-specific molecular network are enriched with multiple AD-related pathways, including cytokine signaling pathways. The proteins from the DAA-specific molecular network are enriched with multiple cytokine signaling pathways.

(C) Inferred cell-cell interactions using CellChat.

(D) Top selected significant ligand-receptor pairs among autoCell-identified cell types. Ligand-receptor interactions were predicted by CellChat (see STAR Methods).

interpolation performance and feature extraction. In addition, we have shown that autoCell can provide greater flexibility in dealing with large datasets different from other imputation algorithms.

Limitations of the study

We acknowledge several potential limitations in autoCell. For example, compared with scVI and SAUCIE, autoCell cannot handle batch effects in scRNA-seq data. However, batch effects are inevitable in studies involving human tissues because the data are usually generated at different times, and batches may affect biological variation. Failure to eliminate batch effects will tarnish downstream analysis and lead to incorrect interpretation of results. In the future, we can improve the model to explicitly allow for the discovery and elimination of batch effects. In addition, with the progress of sequencing technology, multi-omics sequencing technologies such as scMT-seq (single-cell methylome and transcriptome sequencing) and scTrio-seq (single-cell triple omics sequencing) can simultaneously detect DNA methylation and transcriptome data from the same cell. DNA methylation is an important epigenetic regulatory signal. However, its regulatory effect on gene expression in single cells remains a challenge. Therefore, in the future, we will continue to enhance autoCell by implementing a heterogeneous map mosaic to support the integration of single-cell multi-omics data. Finally, compared with methods based on statistical models, the pre-training processing of the autoCell model requires more computing resources, which is more time consuming. Thus, we will study the creation of a more efficient autoCell model through the architecture of block and parallel processing in the near future.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2022.100382.

ACKNOWLEDGMENTS

This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract HHSN261201500003I to R.N. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services nor does mention of trade names, commercial products, or organizations imply endorsement by the US government. This research was supported (in part) by the Intramural Research Program of NIH, National Cancer Institute, Center for Cancer Research.

AUTHOR CONTRIBUTIONS

F.C. conceived the study. X.Z., Junlin Xu, Y.M., C.L., and Jielin Xu performed experiments and data analysis. L.C. and R.N. interpreted the data analysis. X.Z., Junlin Xu, R.N., and F.C. drafted the manuscript and critically revised the manuscript. All authors critically revised and gave final approval of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 26, 2022
Revised: October 31, 2022
Accepted: December 8, 2022
Published: January 5, 2023

REFERENCES

1. Hicks, S.C., Townes, F.W., Teng, M., and Irizarry, R.A. (2018). Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics *19*, 562–578. https://doi.org/10.1093/biostatistics/kxx053.
2. Grün, D., Kester, L., and Van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. Nat. Methods *11*, 637–640. https://doi.org/10.1038/nmeth.2930.
3. Slyper, M., Porter, C.B.M., Ashenberg, O., Waldman, J., Drokhlyansky, E., Wakiro, I., Smillie, C., Smith-Rosario, G., Wu, J., Dionne, D., et al. (2020). A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. Nat. Med. *26*, 792–802. https://doi.org/10.1038/s41591-020-0844-1.
4. Xu, J., Zhang, P., Huang, Y., Zhou, Y., Hou, Y., Bekris, L.M., Lathia, J., Chiang, C.W., Li, L., Pieper, A.A., et al. (2021). Multimodal single-cell/nucleus RNA sequencing data analysis uncovers molecular networks between disease-associated microglia and astrocytes with implications for drug repurposing in Alzheimer's disease. Genome Res. *31*, 1900–1912. https://doi.org/10.1101/gr.272484.120.
5. Wang, H.-Y., Zhao, J.-P., Zheng, C.-H., and Su, Y.S. (2022). scCNC: a method based on capsule network for clustering scRNA-seq data. Bioinformatics *38*, 3703–3709. https://doi.org/10.1093/bioinformatics/btac393.
6. Wang, J., Xia, J., Tan, D., Lin, R., Su, Y., and Zheng, C.H. (2022). scHFC: a hybrid fuzzy clustering method for single-cell RNA-seq data optimized by natural computation. Brief. Bioinform. *23*, bbab588. https://doi.org/10.1093/bib/bbab588.

7. Li, W.V., and Li, J.J. (2018). An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nat. Commun. *9*, 997–999. https://doi.org/10.1038/s41467-018-03405-7.

8. Pierson, E., and Yau, C. (2015). ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. *16*, 241–310. https://doi.org/10.1186/s13059-015-0805-z.

9. Tian, T., Wan, J., Song, Q., and Wei, Z. (2019). Clustering single-cell RNA-seq data with a model-based deep learning approach. Nat. Mach. Intell. *1*, 191–198. https://doi.org/10.1038/s42256-019-0037-0.

10. Wang, D., and Gu, J. (2018). VASC: dimension reduction and visualization of single-cell RNA-seq data by deep variational autoencoder. Dev. Reprod. Biol. *16*, 320–331. https://doi.org/10.1016/j.gpb.2018.08.003.

11. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.-P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. Nat. Commun. *9*, 284–317. https://doi.org/10.1038/s41467-017-02554-5.

12. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., and Batzoglou, S. (2017). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nat. Methods *14*, 414–416. https://doi.org/10.1038/nmeth.4207.

13. Xu, J., Cai, L., Liao, B., Zhu, W., and Yang, J.L. (2021). CMF-Impute: an accurate imputation tool for single-cell RNA-seq data. Bioinformatics *36*, 5563–5564. https://doi.org/10.1093/bioinformatics/btaa664.

14. Prabhakaran, S., Azizi, E., Carr, A., and Pe'er, D. (2016). Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. JMLR Workshop Conf. Proc. *48*, 1070–1079.

15. Arisdakessian, C., Poirion, O., Yunits, B., Zhu, X., and Garmire, L.X. (2019). DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data. Genome Biol. *20*, 211. https://doi.org/10.1186/s13059-019-1837-6.

16. Eraslan, G., Simon, L.M., Mircea, M., Mueller, N.S., and Theis, F.J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. Nat. Commun. *10*, 390–414. https://doi.org/10.1038/s41467-018-07931-2.

17. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat. Methods *15*, 1053–1058. https://doi.org/10.1038/s41592-018-0229-2.

18. Grønbech, C.H., Vording, M.F., Timshel, P.N., Sønderby, C.K., Pers, T.H., and Winther, O. (2020). scVAE: variational auto-encoders for single-cell gene expression data. Bioinformatics *36*, 4415–4422. https://doi.org/10.1093/bioinformatics/btaa293.

19. Wang, J., Ma, A., Chang, Y., Gong, J., Jiang, Y., Qi, R., Wang, C., Fu, H., Ma, Q., and Xu, D. (2021). scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. Nat. Commun. *12*, 1882–1911. https://doi.org/10.1038/s41467-021-22197-x.

20. Zappia, L., Phipson, B., and Oshlack, A. (2017). Splatter: simulation of single-cell RNA sequencing data. Genome Biol. *18*, 174. https://doi.org/10.1186/s13059-017-1305-0.

21. Van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J., Burdziak, C., Moon, K.R., Chaffer, C.L., Pattabiraman, D., et al. (2018). Recovering gene interactions from single-cell data using data diffusion. Cell *174*, 716–729.e27. https://doi.org/10.1016/j.cell.2018.05.061.

22. Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J.I., Raj, A., Li, M., and Zhang, N.R. (2018). SAVER: gene expression recovery for single-cell RNA sequencing. Nat. Methods *15*, 539–542. https://doi.org/10.1038/s41592-018-0033-z.

23. Amodio, M., Van Dijk, D., Srinivasan, K., Chen, W.S., Mohsen, H., Moon, K.R., Campbell, A., Zhao, Y., Wang, X., Venkataswamy, M., et al. (2019). Exploring single-cell data with deep multitasking neural networks. Nat. Methods *16*, 1139–1145. https://doi.org/10.1038/s41592-019-0576-7.

24. Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science *347*, 1138–1142. https://doi.org/10.1126/science.aaa1934.

25. Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187–1201. https://doi.org/10.1016/j.cell.2015.04.044.

26. Petropoulos, S., Edsgärd, D., Reinius, B., Deng, Q., Panula, S.P., Codeluppi, S., Plaza Reyes, A., Linnarsson, S., Sandberg, R., and Lanner, F. (2016). Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. Cell *165*, 1012–1026. https://doi.org/10.1016/j.cell.2016.03.023.

27. Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature *566*, 496–502. https://doi.org/10.1038/s41586-019-0969-x.

28. Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. Nat. Biotechnol. *37*, 547–554. https://doi.org/10.1038/s41587-019-0071-9.

29. Street, K., Risso, D., Fletcher, R.B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genom. *19*, 477. https://doi.org/10.1186/s12864-018-4772-0.

30. Li, X., Wang, K., Lyu, Y., Pan, H., Zhang, J., Stambolian, D., Susztak, K., Reilly, M.P., Hu, G., and Li, M. (2020). Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. Nat. Commun. *11*, 2338. https://doi.org/10.1038/s41467-020-15851-3.

31. Leng, K., Li, E., Eser, R., Piergies, A., Sit, R., Tan, M., Neff, N., Li, S.H., Rodriguez, R.D., Suemoto, C.K., et al. (2021). Molecular characterization of selectively vulnerable neurons in Alzheimer's disease. Nat. Neurosci. *24*, 276–287. https://doi.org/10.1038/s41593-020-00764-7.

32. Cheng, F., Lu, W., Liu, C., Fang, J., Hou, Y., Handy, D.E., Wang, R., Zhao, Y., Yang, Y., Huang, J., et al. (2019). A genome-wide positioning systems network algorithm for in silico drug repurposing. Nat. Commun. *10*, 3476–3514. https://doi.org/10.1038/s41467-019-10744-6.

33. Yamazaki, Y., Zhao, N., Caulfield, T.R., Liu, C.-C., and Bu, G. (2019). Apolipoprotein E and Alzheimer disease: pathobiology and targeting strategies. Nat. Rev. Neurol. *15*, 501–518. https://doi.org/10.1038/s41582-019-0228-7.

34. Congdon, E.E., and Sigurdsson, E.M. (2018). Tau-targeting therapies for Alzheimer disease. Nat. Rev. Neurol. *14*, 399–415. https://doi.org/10.1038/s41582-018-0013-z.

35. Pinner, E., Gruper, Y., Ben Zimra, M., Kristt, D., Laudon, M., Naor, D., and Zisapel, N. (2017). CD44 splice variants as potential players in Alzheimer's disease pathology. J. Alzheimer's Dis. *58*, 1137–1149. https://doi.org/10.3233/JAD-161245.

36. Reichenbach, N., Delekate, A., Plescher, M., Schmitt, F., Krauss, S., Blank, N., Halle, A., and Petzold, G.C. (2019). Inhibition of Stat3-mediated astrogliosis ameliorates pathology in an Alzheimer's disease model. EMBO Mol. Med. *11*, e9665. https://doi.org/10.15252/emmm.201809665.

37. Acosta, C., Anderson, H.D., and Anderson, C.M. (2017). Astrocyte dysfunction in Alzheimer disease. J. Neurosci. Res. *95*, 2430–2447. https://doi.org/10.1002/jnr.24075.

38. González-Reyes, R.E., Nava-Mesa, M.O., Vargas-Sánchez, K., Ariza-Salamanca, D., and Mora-Muñoz, L. (2017). Involvement of astrocytes in alzheimer's disease from a neuroinflammatory and oxidative stress perspective. Front. Mol. Neurosci. *10*, 427. https://doi.org/10.3389/fnmol.2017.00427.

39. Armingol, E., Officer, A., Harismendy, O., and Lewis, N.E. (2021). Deciphering cell–cell interactions and communication from gene expression. Nat. Rev. Genet. *22*, 71–88. https://doi.org/10.1038/s41576-020-00292-x.

40. Wang, K.-S., Xu, N., Wang, L., Aragon, L., Ciubuc, R., Arana, T.B., Mao, C., Petty, L., Briones, D., Su, B.B., et al. (2014). NRG3 gene is associated with

the risk and age at onset of Alzheimer disease. J. Neural. Transm. *121*, 183–192. https://doi.org/10.1007/s00702-013-1091-0.

41. Woo, R.-S., Lee, J.-H., Yu, H.-N., Song, D.-Y., and Baik, T.-K. (2010). Expression of ErbB4 in the apoptotic neurons of Alzheimer's disease brain. Anat. Cell Biol. *43*, 332–339. https://doi.org/10.5115/acb.2010.43.4.332.

42. Chaudhury, A.R., Gerecke, K.M., Wyss, J.M., Morgan, D.G., Gordon, M.N., and Carroll, S.L. (2003). Neuregulin-1 and erbB4 immunoreactivity is associated with neuritic plaques in Alzheimer disease brain and in a transgenic model of Alzheimer disease. J. Neuropathol. Exp. Neurol. *62*, 42–54. https://doi.org/10.1093/jnen/62.1.42.

43. Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., et al. (2017). The human cell atlas. Elife *6*, e27041. https://doi.org/10.7554/eLife.27041.

44. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049–14112. https://doi.org/10.1038/ncomms14049.

45. Romanov, R.A., Zeisel, A., Bakker, J., Girach, F., Hellysaz, A., Tomer, R., Alpár, A., Mulder, J., Clotman, F., Keimpema, E., et al. (2017). Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. Nat. Neurosci. *20*, 176–188. https://doi.org/10.1038/nn.4462.

46. Grubman, A., Chew, G., Ouyang, J.F., Sun, G., Choo, X.Y., McLean, C., Simmons, R.K., Buckberry, S., Vargas-Landin, D.B., Poppe, D., et al. (2019). A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. Nat. Neurosci. *22*, 2087–2097. https://doi.org/10.1038/s41593-019-0539-4.

47. Yang, L., Cheung, N.-M., Li, J., and Fang, J. (2019). Deep clustering by Gaussian mixture variational autoencoders with graph embedding. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 6440–6449. https://doi.org/10.1109/ICCV.2019.00654.

48. Strehl, A., and Ghosh, J. (2002). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. *3*, 583–617.

49. Hubert, L., and Arabie, P. (1985). Comparing partitions. J. Classif. *2*, 193–218. https://doi.org/10.1007/BF01908075.

50. Vinh, N.X., Epps, J., and Bailey, J. (2010). Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. J. Mach. Learn. Res. *11*, 2837–2854.

51. Luck, K., Kim, D.-K., Lambourne, L., Spirohn, K., Begg, B.E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F.J., Charloteaux, B., et al. (2020). A reference map of the human binary protein interactome. Nature *580*, 402–408. https://doi.org/10.1038/s41586-020-2188-x.

52. Cheng, F., Jia, P., Wang, Q., and Zhao, Z. (2014). Quantitative network mapping of the human kinome interactome reveals new clues for rational kinase inhibitor discovery and individualized cancer therapy. Oncotarget *5*, 3697–3710. https://doi.org/10.18632/oncotarget.1984.

53. Peri, S., Navarro, J.D., Kristiansen, T.Z., Amanchy, R., Surendranath, V., Muthusamy, B., Gandhi, T.K.B., Chandrika, K.N., Deshpande, N., Suresh, S., et al. (2004). Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res. *32*, D497–D501. https://doi.org/10.1093/nar/gkh070.

54. Hu, J., Rho, H.-S., Newman, R.H., Zhang, J., Zhu, H., and Qian, J. (2014). PhosphoNetworks: a database for human phosphorylation networks. Bioinformatics *30*, 141–142. https://doi.org/10.1093/bioinformatics/btt627.

55. Hornbeck, P.V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V., and Skrzypek, E. (2015). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res. *43*, D512–D520. https://doi.org/10.1093/nar/gku1267.

56. Dinkel, H., Chica, C., Via, A., Gould, C.M., Jensen, L.J., Gibson, T.J., and Diella, F. (2011). Phospho. ELM: a database of phosphorylation sites—update 2011. Nucleic Acids Res. *39*, D261–D267. https://doi.org/10.1093/nar/gkq1104[.

57. Fazekas, D., Koltai, M., Türei, D., Módos, D., Pálfy, M., Dúl, Z., Zsákai, L., Szalay-Bekő, M., Lenti, K., Farkas, I.J., et al. (2013). SignaLink 2–a signaling pathway resource with multi-layered regulatory networks. BMC Syst. Biol. *7*, 7–15. https://doi.org/10.1186/1752-0509-7-7.

58. Meyer, M.J., Das, J., Wang, X., and Yu, H. (2013). INstruct: a database of high-quality 3D structurally resolved protein interactome networks. Bioinformatics *29*, 1577–1579. https://doi.org/10.1093/bioinformatics/btt181.

59. Huttlin, E.L., Ting, L., Bruckner, R.J., Gebreab, F., Gygi, M.P., Szpyt, J., Tam, S., Zarraga, G., Colby, G., Baltier, K., et al. (2015). The BioPlex network: a systematic exploration of the human interactome. Cell *162*, 425–440. https://doi.org/10.1016/j.cell.2015.06.043.

60. Chatr-Aryamontri, A., Breitkreutz, B.-J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'Donnell, L., et al. (2015). The BioGRID interaction database: 2015 update. Nucleic Acids Res. *43*, D470–D478. https://doi.org/10.1093/nar/gku1204.

61. Cowley, M.J., Pinese, M., Kassahn, K.S., Waddell, N., Pearson, J.V., Grimmond, S.M., Biankin, A.V., Hautaniemi, S., and Wu, J. (2012). PINA v2. 0: mining interactome modules. Nucleic Acids Res. *40*, D862–D865. https://doi.org/10.1093/nar/gkr967.

62. Goel, R., Harsha, H.C., Pandey, A., and Prasad, T.S.K. (2012). Human protein reference database and human Proteinpedia as resources for phosphoproteome analysis. Mol. Biosyst. *8*, 453–463. https://doi.org/10.1039/C1MB05340J.

63. Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A.P., Santonico, E., et al. (2012). MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. *40*, D857–D861. https://doi.org/10.1093/nar/gkr930.

64. Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., Del-Toro, N., et al. (2014). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. *42*, D358–D363. https://doi.org/10.1093/nar/gkt1115.

65. Breuer, K., Foroushani, A.K., Laird, M.R., Chen, C., Sribnaia, A., Lo, R., Winsor, G.L., Hancock, R.E.W., Brinkman, F.S.L., and Lynn, D.J. (2013). InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. Nucleic Acids Res. *41*, D1228–D1233. https://doi.org/10.1093/nar/gks1147.

66. Martens, M., Ammar, A., Riutta, A., Waagmeester, A., Slenter, D.N., Hanspers, K., A Miller, R., Digles, D., Lopes, E.N., Ehrhart, F., et al. (2021). WikiPathways: connecting communities. Nucleic Acids Res. *49*, D613–D621. https://doi.org/10.1093/nar/gkaa1024.

67. Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinf. *14*, 128. https://doi.org/10.1186/1471-2105-14-128.

68. Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Prlic, M., et al. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. *16*, 278. https://doi.org/10.1186/s13059-015-0844-5.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Mouse cortex	Zeisel et al.,[24]	GEO: GSE60361
Embryonic Stem Cells	Klein, A. M. et al.,[25]	GEO: GSE65525
Mouse hypothalamus	Romanov, R.A. et al.,[45]	GEO: GSE74672
Alzheimer's disease	Grubman A et al.,[46]	GEO: GSE138852
Healthy human and principally involves major cell	Zheng et al.,[44]	10X GENOMICS: SRP073767
Fluorescence-activated cell sorting cells in a healthy human	Zheng et al.,[44]	10X GENOMICS: http://support.10xgenomics.com/single-cell/datasets
Human preimplantation embryonic	Petropoulos et al.,[26]	ArrayExpress E-MTAB-3929
Software and algorithms		
MAGIC	Van Dijk, et al.,[21]	https://github.com/KrishnaswamyLab/magic
SAVER	Huang, et al.,[22]	https://github.com/mohuangx/SAVER/releases
DCA	Eraslan et al.,[16]	https://github.com/theislab/dca
scVI	Lopez et al.,[17]	https://github.com/YosefLab/scVI
DESC	Li, et al.,[30]	https://eleozzr.github.io/desc/
scVAE	Grønbech et al.,[18]	https://github.com/scvae/scvae
SAUCIE	Amodio, et al.,[23]	https://github.com/KrishnaswamyLab/SAUCIE/
scGNN	Wang et al.,[19]	https://github.com/juexinwang/scGNN
autoCell	This paper	https://github.com/ChengF-Lab/autoCell, https://doi.org/10.5281/zenodo.7331392.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Feixiong Cheng (chengf@ccf.org).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. The accession numbers for the datasets are listed in the key resources table.
- Source codes are available: https://github.com/ChengF-Lab/autoCell. All original code has been deposited at Zenodo and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Resource Availability: Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

METHOD DETAILS

Deep Gaussian mixture model

Using a Gaussian distribution as the prior probability distribution of z only allows for one mode in the latent representation. In the presence of inherent clustering in data, such as scRNA-seq data where the cells represent different cell types, multiple modes are desirable, for example, one for every cluster or class. This strategy can be implemented by using a Gaussian-mixture model in place of Gaussian distribution.

We aim to cluster a given set of D-dimensional training samples $\{x_n\}_{n=1}^{N}$ into K categories. For each training sample x, we learn a latent feature $z \in R^{M \times 1}$. We assume that the underlying features follow a Gaussian mixture distribution. We introduce a binary vector $c \in \{0,1\}^{K \times 1}$ to indicate which Gaussian component the latent feature z belongs to.

Zero-inflated negative binomial model

ZINB distribution is favored in single-cell RNA data analysis. Previous studies have shown that the ZINB distribution model can effectively denoise single-cell RNA data by modeling highly sparse and over-dispersed data. Therefore, in this study, we use ZINB to model scRNA-seq data:

$$B(x;\mu,\varphi) = \frac{\Gamma(x+\varphi)}{\Gamma(\varphi)}\left(\frac{\varphi}{\varphi+\mu}\right)^{\varphi}\left(\frac{\mu}{\varphi+\mu}\right)^{x}$$

$$ZINB(x;\lambda,\mu,\varphi) = \lambda I_0(x) + (1-\lambda)NB(x;\mu,\varphi) \qquad \text{(Equation 1)}$$

where λ represents the proportion of zero value; μ and φ are the parameters of the negative binomial distribution, and I_0 is the indicator function. When the independent variable is 0, the value is 1, otherwise it is 0.

Generative model

In our model, we assume that the data are drawn from a Gaussian mixture distribution. In particular, for scRNA-Seq data x, we model the generation process as follows:

$$p(c,\pi) = (p_1,\ldots p_k) = \prod_{k=1}^{K} \pi_k^{c_k}$$

$$p(c_k = 1) = \pi_{ik}$$

$$p(z|c_k = 1) = N\left(\mu_k, diag\left(\sigma_k^2\right)\right)$$

$$P_\theta(x|z) = ZINB(x;\lambda,\mu,\varphi) = \lambda I_0(x) + (1-\lambda)NB(x;\mu,\varphi) \qquad \text{(Equation 2)}$$

where c_k and π_k represent the k th entry of c and π, respectively, and π_k must satisfy $\sum_{k=1}^{K}\pi_k = 1$. μ_k and σ_k^2 represent the mean and variance of the k th Gaussian component, respectively. $(\lambda,\mu,\varphi) = g(z_n;\theta)$, where g is a neural network with a parameter θ that can be trained. Detailed parameter information is provided in Table S5.

Inference model

Finding the maximum a posteriori of potential variables and the maximum likelihood estimation of parameters by directly solving the generative model is difficult.[47] In addressing this problem, we use a new distribution $q_\varphi(z,c|x)$ to approximate the posterior distribution $p_\theta(z,c|x)$. The distribution is extracted from a specific category and parameterized by the trainable parameter φ. In particular, we assume that $q_\varphi(z,x|c)$ can be decomposed into $q_\varphi(z,c|x) = q_\varphi(z|x)q_\varphi(x|c)$. Then, we define the following equations:

$$q_\varphi(z|x) = N\left(\tilde{\mu}, diag\left(\tilde{\sigma}^2\right)\right)$$

$$q_\varphi(x|c) = Multinomial(\tilde{\pi}) \qquad \text{(Equation 3)}$$

where

$$\left[\tilde{\mu}, log\left(\tilde{\sigma}^2\right)\right] = f(x;\varphi)$$

$$\tilde{\pi} = f(z;\varphi) \qquad \text{(Equation 4)}$$

here f represents neural networks with parameter φ.

In the framework of the generative and inference models, the parameters can be estimated by maximizing the log likelihood function:

$$\max_{\varphi,\theta} \sum_{i=1}^{N} \ln p_\theta(x_i) \quad \text{(Equation 5)}$$

Equation 5 is usually solved by maximizing the lower limit of evidence of the log likelihood function using multiple parameterization. We observed that the model consists of two networks, namely, g, and f, which were referred to as decoder and encoder, respectively, because they form a variational autoencoder (VAE).

Graph embedding

Graph embedding aims to find low-dimensional features in a sample similarity graph, which keeps the similarity between vertex pairs. In general, under graph embedding, training sample $\{x_n\}$ is regarded as the vertex of similar graph, and the characteristic of similar graph is regarded as the affinity matrix W. The optimal feature $\{z_n^*\}$ can be obtained using Equation 6:

$$\{z_n^*\} = \arg\min_{ZZ^T = I} \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \left|\left|z_i - z_j\right|\right|_2^2 \quad \text{(Equation 6)}$$

where $Z = [z_1, \ldots z_n]$ and w_{ij} denotes the (i,j) th element of the affinity matrix W. The constraint $ZZ^T = I$ is used to avoid trivial solutions. Based on Equation 6, if samples are connected on the graph, then their features will be close to one another. Thus, we hypothesize that if the two samples are connected on the graph, then they should have similar potential features and cluster assignments. Our model considers latent features and cluster assignments as random variables. Therefore, we add a constraint to Equation 6 by measuring the distance of the posterior distributions and obtain the following equation:

$$\max_{\varphi,\theta} \sum_{i=1}^{N} \left(\ln p_\theta(x_i) - \sum_{j=1}^{N} w_{ij} d\left(q_\varphi(z, c|x_i), q_\varphi(z, c|x_j)\right) \right) \quad \text{(Equation 7)}$$

where $d(:,:)$ is a measure of the distance between two distributions. Furthermore, we require $\sum_j w_{ij} = 1$ to balance the weight of each training sample. Here, we select the Jenson–Shannon (JS) divergence:

$$\max_{\varphi,\theta} \sum_{i=1}^{N} \left(\ln p_\theta(x_i) - \sum_{j=1}^{N} w_{ij} JS\left(q_\varphi(z, c|x_i), q_\varphi(z, c|x_j)\right) \right) \quad \text{(Equation 8)}$$

Construction of affinity matrix

In other graph embedding methods, a properly constructed affinity matrix is important. A typical option of affinity matrix is a set of nearest neighbors for a given data point, with a predefined kernel function to calculate their similarity. For example, for a Gaussian kernel, the elements of the affinity matrix are defined as follows:

$$w_{ij} = \begin{cases} \dfrac{1}{a_i} \exp\left(- \dfrac{\left|\left|x_i - x_j\right|\right|_2^2}{2s_i^2} \right) & if x_j \in N(x_i) \\ 0, & othrewise \end{cases} \quad \text{(Equation 9)}$$

where s_i is a predefined scalar; $N(x_i)$ represents the nearest N_s neighbor of set x_i, and the default number is 10. a_i is standardized, that is, $\sum_j w_{ij} = 1$, and the default number is 1.

Imputation evaluation

We evaluated the median L1 distance, cosine similarity, and RMSE between the real expression values of the original dataset and the estimated values obtained by various imputation algorithms. For all flipped entries, X is the row vector of the original expression, and Y is the row vector of its corresponding imputed expression. The L1 distance is the absolute deviation between the values of the original expression and those of the imputed expression. Lower L1 distance indicates higher similarity.

$$L1 = |x - y|, L1 \in [0, +\infty) \quad \text{(Equation 10)}$$

The cosine similarity indicates the dot product between the original expression and estimated expression.

$$\cos(x, y) = \frac{xy^T}{||x||\,||y||}, cos \in [0, 1] \quad \text{(Equation 11)}$$

The RMSE indicates the square root of the quadratic mean of the difference between the original expression and the estimated expression.

$$RMSE(x,y) = \sqrt{\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N}}, RMSE \in [0, +\infty) \qquad \text{(Equation 12)}$$

Evaluation metric for clustering

We selected four commonly used cluster evaluation indicators, including normalized mutual information (NMI),[48] adjusted Rand index (ARI),[49] completeness (COM),[50] and homogeneity (HOM).[50]

Adjusted rand index

ARI is an improved version of the rand coefficient. In general, the rand coefficient is used to evaluate the clustering results by calculating the similarity between the two clusters. The adjustment of the rand coefficient is an improvement of the rand coefficient based on probability regularization. We define a confusion matrix. In the case of a given gold standard, *a* is the number of cell pairs correctly classified into the same category by clustering; *b* is the number of cell pairs that belong to different categories but are divided into the same cluster; *c* is the number of cell pairs that belong to the same category but are divided into the number of different clusters, and *d* is the number of cell pairs correctly divided into different clusters.

$$ARI = \frac{2(ad - bc)}{(a+b)(b+d)+(a+c)(c+d)}. \qquad \text{(Equation 13)}$$

Normalized mutual information

Mutual information (MI) measures the degree of agreement between two dataset distributions. MI is also an important metric of information, which refers to the correlation between two sets of events. Moreover, NMI is improved on the basis of mutual information. $U = \{u_1, u_2, u_3 \ldots, u_n\}$ and $V = \{v_1, v_2, v_3 \ldots, v_n\}$ denote the gold standard and partition obtained by k-means clustering algorithm, respectively.

$$\text{NMI} = \frac{2I(U,V)}{H(U)+H(V)} \qquad \text{(Equation 14)}$$

$$I(U,V) = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{|u_i \cap v_j|}{N}\log\frac{N|u_i \cap v_j|}{|u_i| \times |v_j|} \qquad \text{(Equation 15)}$$

$$H(U) = -\sum_{i=1}^{n}\frac{u_i}{N}\log\frac{u_i}{N} \text{ and } H(V) = -\sum_{j=1}^{n}\frac{v_j}{N}\log\frac{v_j}{N} \qquad \text{(Equation 16)}$$

Completeness

COM indicates that samples of the same category are classified into the same cluster. If all samples of the same type are grouped in the same cluster, then the integrity is 1. If the samples of the same type are grouped in different clusters, then the conditional empirical entropy $H(V|U)$ is calculated. The larger the value, the smaller the completeness.

$$COM = 1 - \frac{H(V|U)}{H(V)} \qquad \text{(Equation 17)}$$

Homogeneity

HOM indicates that each cluster contains only a single category of samples. If only one category is found in a cluster, then HOM is 1. If multiple categories are found, then the conditional empirical entropy $H(U|V)$ of the cluster under the category is calculated. The larger the value, the smaller the homogeneity.

$$HOM = 1 - \frac{H(U|V)}{H(U)} \qquad \text{(Equation 18)}$$

For all metrics, including ARI, NMI, COM, and HOM, a larger value (up to 1) indicated good performance.

Pseudotime order score (POS)

In measuring the accuracy of reconstruction pseudotime, we defined a POS as follows: $POS = C/(N_c + C)$, where C and N_c represent the number of cell pairs that are consistent and inconsistent between the inferred pseudotime and the gold standard (e.g., true data collection time), respectively.

Protein-protein interactome (PPI) network

To build the comprehensive human interactome from the most contemporary data available, we assembled 18 commonly used PPI databases with experimental evidence and the in-house systematic human PPI that we have previously utilized: (i) binary PPIs tested by high-throughput yeast-two-hybrid (Y2H) system;[51] (ii) kinase-substrate interactions by literature-derived low-throughput and high-throughput experiments from KinomeNetworkX,[52] Human Protein Resource Database (HPRD),[53] PhosphoNetworks,[54] PhosphositePlus,[55] DbPTM 3.0 and Phospho.ELM;[56] (iii) signaling networks by literature-derived low-throughput experiments from the SignaLink2.0;[57] (iv) binary PPIs from three-dimensional protein structures from Instruct;[58] (v) protein complexes data (∼56,000 candidate interactions) identified by a robust affinity purification-mass spectrometry collected from BioPlex V2.0;[59] and (vi) carefully literature-curated PPIs identified by affinity purification followed by mass spectrometry from BioGRID,[60] PINA,[61] HPRD,[62] MINT,[63] IntAct,[64] and InnateDB.[65] Herein, the human interactome constructed in this way includes 351,444 PPIs connecting 17,706 unique human proteins.

Description of GPSnet

GPSnet[32] algorithms take two inputs: node score and one background PPI network. The node score was defined as |log2FC| for significant (FDR <0.05) differentially expressed genes (DEGs), and 0 for non-significant DEGs. GPSnet builds raw module in each iteration by starting with a random seed gene/protein (node). After that, one of the candidate neighboring genes that satisfy the subsequent two conditions is added: (1) connectivity significance P(i) (Equation 19) is less than 0.01; (2) the updated module score is greater than the current one (Equation 20). We repeated steps (1) and (2) until no more genes (nodes) can be added to each raw module. In this study, we built ∼100,000 raw modules ranked by module scores with the corresponding module score computed in (Equation 20). All generated raw modules are ranked in decreasing module score order. The final network module is generated by assembling truncated top-ranked raw modules.

$$P(i) = \sum_{d=d_n}^{d_i} \frac{\binom{n}{d}\binom{N-n}{d_i-d}}{\binom{N}{d_i}} \qquad \text{(Equation 19)}$$

$$MS_{n+i}(i) = \frac{(s(i) - \mu) + \sum_{j \in M}(S(j) - \mu)}{\sqrt{n+1}} \qquad \text{(Equation 20)}$$

Where, N denotes all proteins/genes in the PPI, n represents numbers of nodes in the module, d_n is the numbers of neighbors of node i in the current raw module, d_i is the degree of gene i, $MS_{n+1}(i)$ denotes the updated module score if adding node i to current raw module, $s(i)$ denotes the score of node i, M denotes the current module, and μ is the average node score of all genes in the complete PPI network.

Datasets and pre-processing

autoCell takes raw sc/snRNA-seq gene expression profile as input. Data filtering and quality control are the first step of data pre-processing. Given the high loss rate of sc/snRNA-seq expression data, only genes expressed as non-zero in more than 1% of cells and cells expressed as non-zero in more than 1% of genes are included. Then, the genes are sorted on the basis of the SD that is, the first k genes (Table S6) in the variance are used. In addition, our model automatically log-transforms all data.

In order to test the performance of autoCell, we analyzed two simulated datasets and seven commonly used real-world datasets (Table S6) from both human and mouse to test performance of autoCell across cross-species single-cell/nuclei RNA-seq datasets. Real datasets were obtained from diverse sequencing platforms and experimental protocols (Table S6).

The simulation dataset was generated using the splatter R package. For the two sets of simulated datasets, the following parameters were used in the setParams() function: batchCells = 500, nGenes = 2000, group.prob = c(0.30, 0.3, 0.4), de.prob = c(0.05, 0.08, 0.01), de.facLoc = 0.5, de.facScale = 0.8. The following parameters were used in the splatSimulate() R function: dropout.shape = c(−0.20, −0.20, −0.20) or dropout.shape = c(−0.05, −0.05, −0.05), dropout.mid = c(0,0,0), dropout.type = "group".

The Zeisel[24] dataset consists of 3,005 cells from the somatosensory cortex and hippocampus of the mouse brain (Table S6). The Zeisel dataset has real labels for seven different cell types, including pyramidal cells, oligodendrocytes, parietal cells, interneurons, astrocytes, ependymal cells, and endothelial cells in the brain.

The Klein dataset[25] consists of 2717 cells derived from mouse embryonic stem cells (Table S6). This dataset reveals the heterogeneity of the population structure and differentiation after the withdrawal of leukemia inhibitory factor. The reproducibility of these high-throughput single-cell data allows people to deconstruct cell populations and infer gene expression relationships.

The Romanov dataset[45] consists of 2881 cells from the mouse hypothalamus (Table S6). The dataset includes 1001 oligodendrocytes, 267 astrocytes, 356 ependymal cells, 48 microglia, 240 endothelial cells, 71 vascular smooth muscle cells, and 888 neurons.

The AD dataset (GEO accession number GSE138852)[46] contains 13,214 single-core scRNA-Seq datasets collected from six AD and six control brains (Table S6). This dataset has true labels for eight different cell types: microglia, astrocyte neuron, oligodendrocyte, OPC, endothelial, unidentified, and hybrid.

The Zheng-68k[44] dataset contains fresh peripheral blood mononuclear cells from healthy people, primarily involving the main cell types of peripheral blood mononuclear cells, such as T-cells (T), Natural killer (NK) cells, B-cells (B), and myeloid cells (Table S6). In addition, the Zheng-73k[44] dataset is composed of fluorescently activated sorting cells from healthy people, including T, NK, and B cells.

The Petropoulos dataset[26] included the single cells from five stages of human preimplantation embryonic development from the developmental day (E) 3 to day 7(Table S6).

QUANTIFICATION AND STATISTICAL ANALYSIS

Enrichment analysis

The pathway enrichment analyses were conducted using WikiPathways[66] from Enrichr.[67]

Differential expression analyses

With the cell types annotated by autoCell, we utilized one R package ‘MAST’[68] for the sequential differential expression analyses. The results of differential expression analyses were used as input for the GPSnet algorithm (Method details - Description of GPSnet).