nature biotechnology

Brief Communication

https://doi.org/10.1038/s41587-023-01772-1

scDesign3 generates realistic in silico data for multimodal single-cell and spatial omics

Received: 20 September 2022

Accepted: 30 March 2023

Published online: 11 May 2023

Check for updates

Dongyuan Song[1], Qingyang Wang[2], Guanao Yan[2], Tianyang Liu[2], Tianyi Sun[2] & Jingyi Jessica Li[1,2,3,4,5,6]

We present a statistical simulator, scDesign3, to generate realistic single-cell and spatial omics data, including various cell states, experimental designs and feature modalities, by learning interpretable parameters from real data. Using a unified probabilistic model for single-cell and spatial omics data, scDesign3 infers biologically meaningful parameters; assesses the goodness-of-fit of inferred cell clusters, trajectories and spatial locations; and generates in silico negative and positive controls for benchmarking computational tools.

Single-cell and spatial omics technologies provided unprecedented multimodal views of individual cells. First, single-cell RNA sequencing (scRNA-seq) was developed to measure cells' transcriptomes, enabling the discovery of discrete cell types and continuous cell trajectories[1,2]. Later, other single-cell omics technologies were developed to measure additional molecular feature modalities, including chromatin accessibility[3,4], DNA methylation[5] and protein abundance[6]. More recently, single-cell multiomics technologies were invented to measure more than one feature modality simultaneously[7,8]. In parallel to single-cell omics, spatial transcriptomics technologies were advanced to profile transcriptomes with cells' spatial locations recorded[9–12].

Thousands of computational methods have been developed for various tasks[13], making method benchmarking a pressing challenge. Fair benchmarking demands in silico data that contain ground truths and mimic real data, thus calling for realistic simulators. Two benchmark studies of simulators[14,15] found that reference-based scRNA-seq simulators, which require training on real data, were more realistic than de novo simulators, which use preset theoretical models[15]. The two studies also found that, although some reference-based simulators[16–18] generated realistic scRNA-seq data from discrete cell types[14,15], few reference-based simulators could generate data from continuous cell trajectories[15,19–22]. Moreover, realistic simulators were lacking for single-cell omics other than scRNA-seq[23], not to mention single-cell multiomics and spatial transcriptomics (see Supplementary Methods for discussion on recent advances). Hence, a large gap existed between the diverse benchmarking needs and the limited functionalities of existing simulators.

To fill in the gap, here, we introduce scDesign3, a simulator that generates realistic synthetic data from diverse settings, including cell latent structures, feature modalities, spatial locations and experimental designs (Fig. 1a). Supplementary Table 1 lists a detailed comparison of scDesign3 with the previous two versions, scDesign[24] and scDesign2[16]. scDesign3 offers a probabilistic model that unifies the generation and inference for single-cell and spatial omics data. The model's interpretable parameters and likelihood enable scDesign3 to generate customized in silico data and unsupervisedly assess the goodness-of-fit of inferred cell latent structures (for example, clusters, trajectories and spatial locations) (Fig. 2a).

As an overview, we verified scDesign3's two functionalities—simulation and interpretation—sequentially. First, we show that the scDesign3 model is reasonable in that its synthetic data well mimic real data given high-quality cell-type labels and cell trajectories. Second, assuming the scDesign3 model is reasonable, we show that scDesign3 allows model-based interpretation of real data, including assessment of the goodness-of-fit of inferred cell latent structures.

scDesign3 functionality 1 (simulation)

We verified scDesign3 as a realistic and versatile simulator in four exemplar settings: (1) scRNA-seq of continuous cell trajectories, (2) spatial transcriptomics, (3) single-cell epigenomics and (4) single-cell multiomics (Fig. 1). We show that the synthetic data of scDesign3 resembled the left-out test data consistently.

In the first setting, scDesign3 mimicked three scRNA-seq datasets containing single or bifurcating cell trajectories (EMBRYO, MARROW

[1]Bioinformatics Interdepartmental Ph.D. Program, University of California, Los Angeles, CA, USA. [2]Department of Statistics, University of California, Los Angeles, CA, USA. [3]Department of Human Genetics, University of California, Los Angeles, CA, USA. [4]Department of Computational Medicine, University of California, Los Angeles, CA, USA. [5]Department of Biostatistics, University of California, Los Angeles, CA, USA. [6]Radcliffe Institute for Advanced Study, Harvard University, Cambridge, MA, USA. e-mail: jli@stat.ucla.edu

and PANCREAS in Supplementary Table 2). Figure 1b–c, Extended Data Figs. 1 and 2c,d and Supplementary Fig. 1c,d show that scDesign3 generated realistic synthetic cells that resembled left-out real cells, reflected by high mean local inverse Simpson's index (mLISI) values[25]. Moreover, scDesign3 preserved eight gene- and cell-specific characteristics described in Methods (Extended Data Figs. 1 and 2a,b and Supplementary Fig. 1a,b). Due to the lack of reference-based simulators for continuous cell trajectories, we benchmarked scDesign3 against ZINB-WaVE, muscat and SPARSIM—three top-performing simulators for discrete cell types[14,15]—and a deep-learning-based simulator, scGAN[26]. scDesign3 outperformed these simulators in generating more realistic synthetic cells and in better preserving the gene- and cell-specific characteristics, especially cell–cell distances and gene–gene correlations (Fig. 1b,c, Extended Data Figs. 1 and 2 and Supplementary Fig. 1).

In the second setting, scDesign3 emulated four spatial transcriptomics datasets generated by the 10x Visium and Slide-seq technologies (VISIUM, SLIDE, OVARIAN and ACINAR in Supplementary Table 2). First, Fig. 1d,e and Extended Data Fig. 3 show that scDesign3 recapitulated the expression patterns of spatially variable genes. Second, Extended Data Fig. 4a,b and Supplementary Figs. 2, 3 and 4a,b show that scDesign3 preserved the eight gene- and cell-specific characteristics. Third, Extended Data Fig. 4c,d and Supplementary Figs. 2, 3 and 4c,d use two-dimensional cell embeddings to confirm that the synthetic data of scDesign3 resembled the test data. Fourth, scDesign3 mimicked spatial transcriptomics data so that three prediction algorithms had highly consistent performance when trained on real data or scDesign3 synthetic data (Extended Data Fig. 5). Fifth, the scDesign3 model adapted to complex spatial patterns in less-structured cancer tissues (Extended Data Fig. 6). Sixth, given a pair of scRNA-seq data and spot-resolution spatial transcriptomics data (where each spot contains multiple cells), scDesign3 can generate realistic spot-resolution spatial transcriptomics data with cell-type proportions specified at each spot (Fig. 1f and Extended Data Fig. 7a). Using this functionality to benchmark cell-type deconvolution algorithms for spatial transcriptomics data, we had consistent results with a benchmark study[27] that CARD[27] and RCTD[28] outperformed SPOTlight[29] in estimating cell-type proportions, though we also found that the three algorithms performed similarly well in estimating each cell type's relative abundance distribution across the spots (Extended Data Fig. 7b).

In the third setting, scDesign3 resembled two single-cell chromatin accessibility datasets profiled by the 10x single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq) and single-cell combinatorial indexing assay for transposase-accessible chromatin using sequencing (sci-ATAC-seq) protocols (ATAC and SCIATAC in Supplementary Table 2). For both protocols, scDesign3 generated synthetic cells whose read counts in peak regions resembled those of real cells (Figs. 1g and 1h, left, Extended Data Fig. 8 and Supplementary Fig. 5). Moreover, coupled with our newly developed read simulator scReadSim[30], scDesign3 enabled the generation of realistic synthetic reads, unblocking the capacity for benchmarking read-level bioinformatics tools (Fig. 1h, right).

In the fourth setting, scDesign3 mimicked a cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) dataset (CITE in Supplementary Table 2) and simulated a multiomics dataset from 'separately' measured RNA expression and DNA methylation modalities (SCGEM in Supplementary Table 2). First, scDesign3 resembled the CITE-seq dataset by simultaneously simulating the expression levels of genes and surface proteins (Extended Data Fig. 9a,c,d). Figure 1i shows that the RNA and protein expression levels of three exemplary surface proteins are highly consistent between the synthetic data and the test data. Moreover, scDesign3 recapitulated the correlations between RNA and protein expression levels (Extended Data Fig. 9b). Second, scDesign3 simulated a single-cell multiomics dataset with joint RNA expression and DNA methylation modalities by learning from two single-omics datasets (Fig. 1j, left) with joint low-dimensional cell embeddings found by Pamona[31]. This synthetic multiomics dataset preserved the cell trajectory in the two single-omics datasets (Fig. 1j, right). The functionality to generate multiomics data from single-omics data allows scDesign3 to benchmark the computational methods that integrate modalities from unmatched cells[32].

scDesign3 functionality 2 (interpretation)

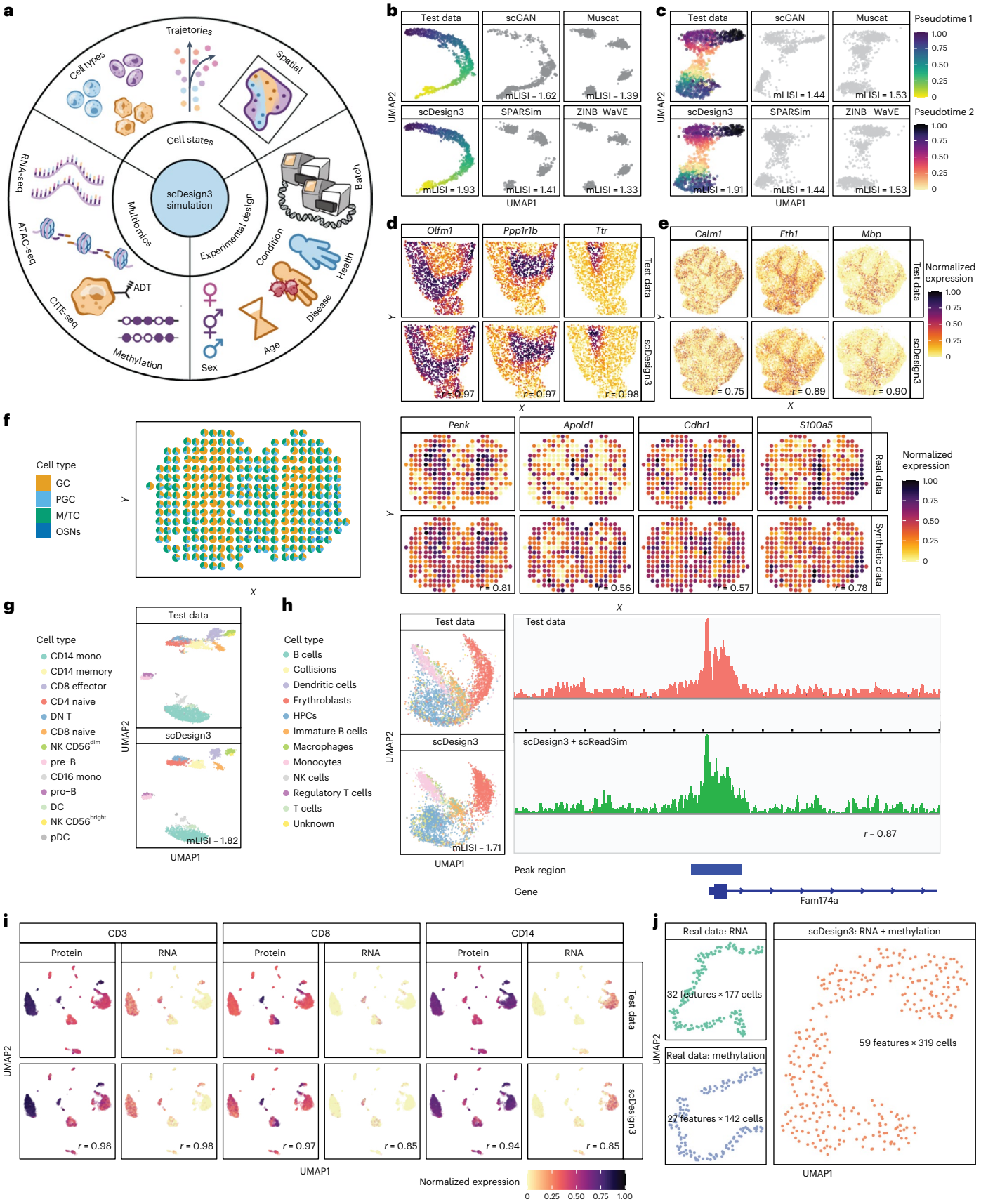
Providing a universal probabilistic model for single-cell and spatial omics data, scDesign3 has broad applications beyond generating synthetic data. We investigated three prominent applications of the scDesign3 model: model parameters, model selection and model alteration (Fig. 2a).

First, the scDesign3 model has an interpretable parametric structure consisting of genes' marginal distributional parameters and pairwise gene correlations. Moreover, the scDesign3 model is flexible to incorporate diverse cell covariates via the use of generalized additive models (GAMs) and Gaussian process (Methods), which allow the estimation of nonlinear gene expression changes along cell trajectories (Fig. 2b) and across spatial locations (Fig. 2c). Besides inferring individual genes' expression characteristics, scDesign3 also estimates pairwise gene correlations conditional on cell covariates, thus providing insights into potential gene regulatory relationships. Specifically, scDesign3 estimates gene correlations by two statistical techniques, Gaussian copula and vine copula, which have complementary advantages (Methods): Gaussian copula is fast but outputs only a gene correlation matrix; vine copula is slow but interpretable by outputting a gene 'vine' with the top layer indicating the most highly correlated genes (that is, 'hub genes'). Applied to an scRNA-seq dataset of human peripheral blood mononuclear cells with four cell types (ZHENGMIX4

Fig. 1 | scDesign3 generates realistic synthetic data of diverse single-cell and spatial omics technologies. a, An overview of scDesign3's simulation functionalities: cell states (for example, discrete cell types, continuous trajectories and spatial locations); multiomics modalities (for example, RNA sequencing (RNA-seq), ATAC-seq, CITE-seq and methylation); and experimental designs (for example, batches, conditions, sex and age). ADT, antibody derived tag. **b**,**c**, scDesign3 outperformed existing simulators scGAN, muscat, SPARSim and ZINB-WaVE in simulating scRNA-seq datasets with a single trajectory (**b**) and bifurcating trajectories (**c**). Larger mLISI values represent better resemblance between synthetic data and test data. **d**,**e**, scDesign3 simulated realistic gene expression patterns in spatial transcriptomics datasets measured by 10x Visium (**d**) and Slide-seq (**e**). Large Pearson correlation coefficients (*r*) represent similar spatial patterns in synthetic data and test data. **f**, Using paired scRNA-seq data and spatial transcriptomics data (MOB-SC and MOB-SP in Supplementary Table 2) as input, we defined the 'ground truth' cell-type proportions at each spot (left), with the cell types including granule cells (GC), periglomerular cells (PGC), mitral/tufted cells (M/TC) and olfactory sensory neurons (OSNs). Each color represents a cell type. With the cell-type proportions, scDesign3 generated synthetic spatial transcriptomics data in which every spot is a mixture of synthetic single cells, given the spot's cell-type proportions. The four cell-type marker genes exhibit similar spatial expression patterns in real data (right top) and synthetic data (right bottom). Large *r* values represent similar expression patterns in synthetic data and test data. **g**, scDesign3 simulated a realistic scATAC-seq dataset at the count level. DC, dendritic cells; DN T, double-negative T cells; mono, monocytes; NK, natural killer cells; pDC, plasmacytoid dendritic cells. **h**, scDesign3 simulated a realistic sci-ATAC-seq dataset at both the count level (left, Uniform Manifold Approximation and Projection (UMAP) visualizations of real and synthetic cells based on peak counts) and the read level when coupled with scReadSim[30] (right, pseudobulk read coverages). HPCs, hematopoietic progenitor cells. **i**, scDesign3 simulated realistic CITE-seq data. Three genes' protein and RNA abundances are shown on the cell UMAP embeddings in test data (top) and synthetic data (bottom). Large *r* values represent similar expression patterns in synthetic data and test data. **j**, scDesign3 generated a multiomics (RNA expression + DNA methylation) dataset (right) by learning from two real single-omics datasets with RNA expression or DNA methylation only (left). The synthetic data preserved the linear cell topology.

https://doi.org/10.1038/s41587-023-01772-1

in Supplementary Table 2), Gaussian copula revealed similar gene correlation matrices for similar cell types (regulatory T cells versus naive cytotoxic T cells) and distinct gene correlation matrices for distinct cell types (CD14$^+$ monocytes versus naive cytotoxic T cells) (Fig. 2d, top); vine copula discovered canonical cell-type marker genes as hub genes: *LYZ* for CD14$^+$ monocytes and *CD79A* for B cells (Fig. 2d, bottom).



https://doi.org/10.1038/s41587-023-01772-1

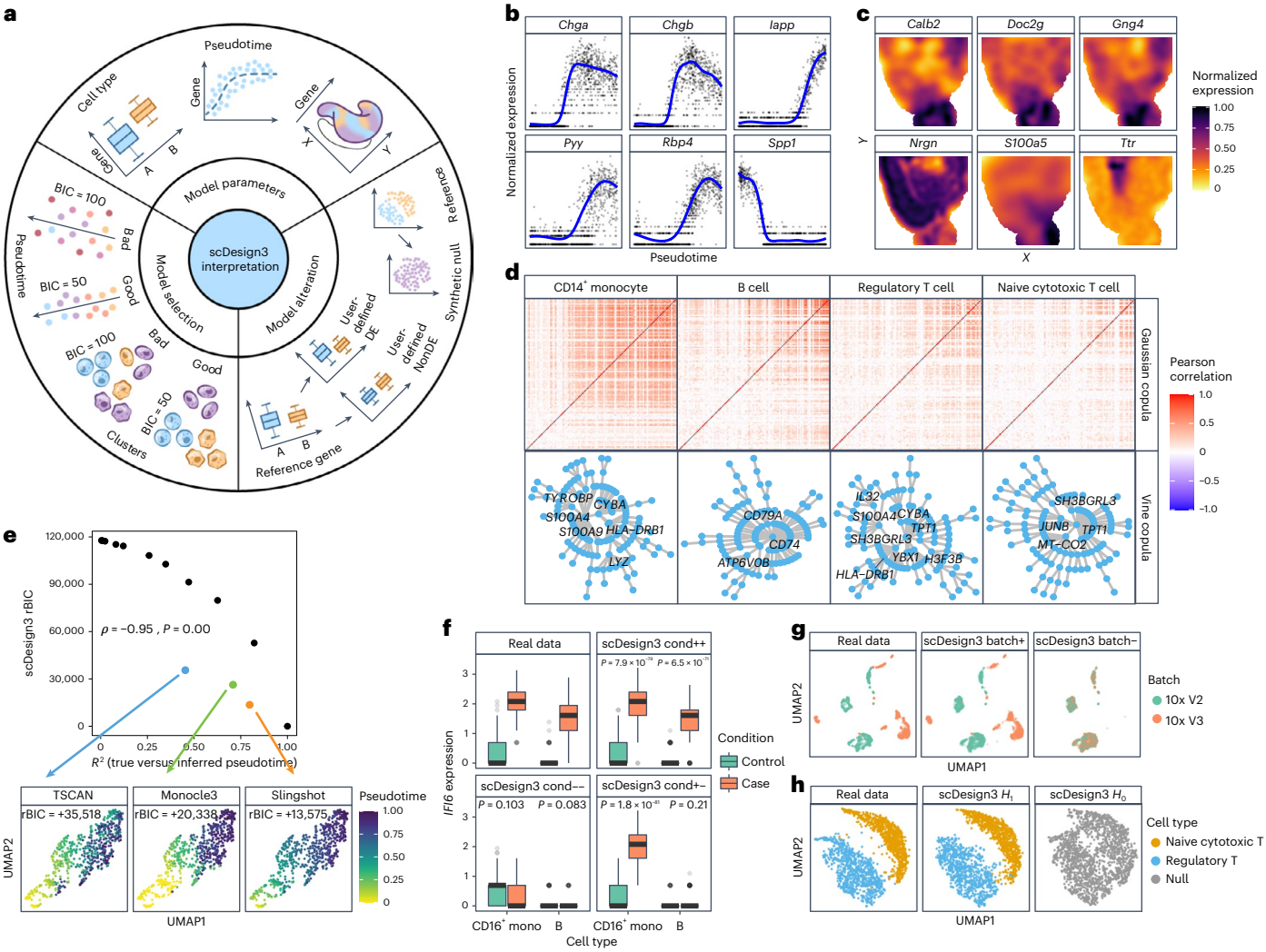


Fig. 2 | scDesign3 enables comprehensive interpretation of real data. **a**, Summary of scDesign3's interpretation functionalities. DE, differentially expressed; NonDE, non-DE. **b**, scDesign3 estimated six genes' expression trends along cell pseudotime (PANCREAS in Supplementary Table 2). **c**, scDesign3 estimated six genes' spatial expression trends (VISIUM in Supplementary Table 2). **d**, scDesign3 estimated cell-type-specific gene correlations (ZHENGMIX4 in Supplementary Table 2): correlation matrices by Gaussian copula (top); vine representations by vine copula (bottom), with genes in the first layer (roughly the genes strongly correlated) labeled. **e**, scDesign3's unsupervised assessment of goodness-of-fit. On synthetic scRNA-seq data with true pseudotimes (based on EMBYRO in Supplementary Table 2), the scDesign3 BIC and the supervised R^2 were evaluated on inferred pseudotimes of TSCAN (blue), Monocle3 (green) and Slingshot (orange), with perturbed true pseudotimes (black) as reference. Top, relative BIC (rBIC = BIC minus the smallest BIC) versus R^2; the P value (P) is from the one-sided test of Spearman's rank correlation ρ (H_0: $\rho = 0$; H_1: $\rho < 0$). Bottom, UMAP visualization of the three methods' inferred pseudotimes. **f**, In the CONDITION dataset (Supplementary Table 2), gene *IFI6* was up-regulated in both CD16$^+$ monocytes and B cells from control (green) to stimulation (red). scDesign3 simulated data where *IFI6* was up-regulated in both cell types (cond++), unchanged in both cell types (cond−−) or up-regulated in CD16$^+$ monocytes only (cond+−). The box center lines, bounds and whiskers denote the medians, first and third quartiles, and minimum and maximum values within 1.5 × the interquartile range of the box limits, respectively (the control and stimulation conditions have $n_{control} = 1{,}772$ and $n_{stimulation} = 2{,}188$ cells, respectively). The P values (P) are from the two-sided Wilcoxon rank-sum test. **g**, The BATCH dataset (Supplementary Table 2) contains two batches (left), which were measured by 10x Chromium Version 2 and Version 3 (10x V2 and 10x V3), respectively. scDesign3 preserved the batch effects in synthetic data generation (batch+) or generated synthetic data without batch effects (batch−). **h**, The ZHENGMIX4 dataset (Supplementary Table 2) contains two cell types (left). scDesign3 resembled the real data under the alternative hypothesis (H_1: two cell types existed) (middle) or generated synthetic data under the null hypothesis (H_0: one cell type existed) (right).

Second, scDesign3 embraces likelihood-based model selection criteria such as Akaike information criterion (AIC) and Bayesian information criterion (BIC), allowing scDesign3 to evaluate the 'goodness-of-fit' of a model to data and to compare competing models. A noteworthy application is evaluating how inferred cell latent structures (clusters, trajectories and spatial locations) describe data, that is, assessing latent structures from the goodness-of-fit perspective without ground truths or external knowledge. Although the scDesign3 model does not represent ground truths, we demonstrated that scDesign3 AIC and BIC are useful 'unsupervised' criteria for assessing how well latent structures agree with data under the scDesign3 model.

For cell clustering, we benchmarked scDesign3 BIC against the 'supervised' adjusted Rand index (ARI) (Methods) and the newly proposed 'unsupervised' clustering deviation index[33] on eight datasets with known cell types[34]. The results show that scDesign3 BIC agreed well with ARI (mean Spearman correlation < −0.7) and had better or similar performance compared with clustering deviation index (Extended Data Fig. 10b). For pseudotime inference, scDesign3 BIC correlated well (mean Spearman correlation < −0.7) with the 'supervised' R^2 (Methods) on multiple synthetic datasets with true pseudotimes (Fig. 2e, top, and Extended Data Fig. 10a). Applied to three pseudotime inference methods, scDesign3 BIC found the

pseudotimes inferred by Slingshot[35] agreed better with data (smaller BIC) than those inferred by TSCAN[36] and Monocle3[2] (Fig. 2e, bottom). For spatial location inference, we found scDesign3 AIC correlated well (mean Spearman correlation < −0.7) with the 'supervised' mean cosine similarity (Methods) on two spatial transcriptomics datasets (MOUSE-CORTEX and MOUSE-VISUAL in Supplementary Table 2), suggesting that scDesign3 AIC is effective in assessing spatial locations' goodness-of-fit (Extended Data Fig. 10c). Note that scDesign3 AIC outperformed BIC in this case, possibly because AIC prefers more complex models, which can better fit complex spatial data.

Third, scDesign3 has a model alteration functionality: given the scDesign3 model parameters estimated on real data, users can alter these parameters to reflect a hypothesis and generate the corresponding in silico data with real data characteristics. This functionality makes scDesign3 advantageous over deep-learning-based simulators[26], which cannot be easily altered to reflect a hypothesis. We demonstrated how to use this functionality in three examples. First, scDesign3 can generate synthetic data with different cell-type-specific condition effects (Fig. 2f). In a real dataset (CONDITION in Supplementary Table 2), gene *IFI6*'s expression was up-regulated after stimulation in both $CD16^+$ monocytes and B cells (Fig. 2f, top-left). With scDesign3's fitted model, we altered *IFI6*'s mean parameters to make *IFI6*'s expression up-regulated (Fig. 2f, top-right) or unchanged (Fig. 2f, bottom-left) in both cell types, or up-regulated in $CD16^+$ monocytes only (Fig. 2f, bottom-right). Second, scDesign3 can generate synthetic data with or without batch effects (Fig. 2g). Trained on a real dataset (BATCH in Supplementary Table 2) containing two batches (Fig. 2g, left), scDesign3 generated synthetic data retaining the batch effects (Fig. 2g, middle); then we altered the batch parameter in the fitted scDesign3 model to generate synthetic data without batch effects (Fig. 2g right). Third, scDesign3 can generate synthetic data under the null hypothesis (H_0) that only one cell type exists and the alternative hypothesis (H_1) that two cell types exist (Fig. 2h). Given a real dataset (ZHENGMIX4 in Supplementary Table 2 and Fig. 2h, left), under H_1, we fitted the model using cell-type labels (Fig. 2h, middle); under H_0, we fitted the model by assuming all cells are of one type (Fig. 2h, right). Using the two fitted models, scDesign3 generated synthetic data under H_1 and H_0. Particularly, the synthetic data under H_0 can serve as the in silico negative control for benchmarking cell-type identification methods.

In summary, scDesign3 accommodates various cell statuses, diverse omics modalities and complex experimental designs. Although the scDesign3 model should not be treated as the true model, its interpretable parameters precede functionalities besides data simulation. First, scDesign3 model parameters offer a comprehensive interpretation of real data. Second, scDesign3 allows likelihood-based model selection to assess the goodness-of-fit of inferred cell clusters, trajectories and spatial locations. Of course, this unsupervised model-based assessment cannot replace supervised metrics or compare models with different types of cell latent structures (for example, cell clusters versus trajectories). Third, scDesign3 can generate synthetic data under specific hypotheses by having its model parameters altered.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-023-01772-1.

References

1. Tang, F. et al. mRNA-seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
2. Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
3. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
4. Cusanovich, D. A. et al. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
5. Karemaker, I. D. & Vermeulen, M. Single-cell DNA methylation profiling: technologies and biological applications. *Trends Biotechnol.* **36**, 952–965 (2018).
6. Bendall, S. C. et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332**, 687–696 (2011).
7. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* **37**, 1452–1457 (2019).
8. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
9. Rao, N., Clark, S. & Habern, O. Bridging genomics and tissue pathology: 10x genomics explores new frontiers with the visium spatial gene expression solution. *Genet. Eng. Biotechnol. News* **40**, 50–51 (2020).
10. Rodriques, S. G. et al. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 (2019).
11. Stickels, R. R. et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* **39**, 313–319 (2021).
12. Moffitt, J. R. et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, eaau5324 (2018).
13. Efremova, M. & Teichmann, S. A. Computational methods for single-cell omics across modalities. *Nat. Methods* **17**, 14–17 (2020).
14. Cao, Y., Yang, P. & Yang, J. Y. H. A benchmark study of simulation methods for single-cell RNA sequencing data. *Nat. Commun.* **12**, 6911 (2021).
15. Crowell, H. L., Morillo Leonardo, S. X., Soneson, C. & Robinson, M. D. The shaky foundations of simulating single-cell RNA sequencing data. *Genome Biol.* **24**, 62 (2023).
16. Sun, T., Song, D., Li, W. V. & Li, J. J. scDesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured. *Genome Biol.* **22**, 163 (2021).
17. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J.-P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* **9**, 284 (2018).
18. Crowell, H. L. et al. Muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nat. Commun.* **11**, 6077 (2020).
19. Cannoodt, R., Saelens, W., Deconinck, L. & Saeys, Y. Spearheading future omics analyses using dyngen, a multi-modal simulator of single cells. *Nat. Commun.* **12**, 3942 (2021).
20. Dibaeinia, P. & Sinha, S. Sergio: a single-cell expression simulator guided by gene regulatory networks. *Cell Syst.* **11**, 252–271 (2020).
21. Papadopoulos, N., Gonzalo, P. R. & Söding, J. Prosstt: probabilistic simulation of single-cell RNA-seq data for complex differentiation processes. *Bioinformatics* **35**, 3517–3519 (2019).
22. Tian, J., Wang, J. & Roeder, K. Esco: single cell expression simulation incorporating gene co-expression. *Bioinformatics* **37**, 2374–2381 (2021).
23. Navidi, Z., Zhang, L. & Wang, B. simATAC: a single-cell ATAC-seq simulation framework. *Genome Biol.* **22**, 74 (2021).
24. Li, W. V. & Li, J. J. A statistical simulator scDesign for rational scRNA-seq experimental design. *Bioinformatics* **35**, i41–i50 (2019).

25. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
26. Marouf, M. et al. Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. *Nat. Commun.* **11**, 166 (2020).
27. Ma, Y. & Zhou, X. Spatially informed cell-type deconvolution for spatial transcriptomics. *Nat. Biotechnol.* **40**, 1349–1359 (2022).
28. Cable, D. M. et al. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat. Biotechnol.* **40**, 517–526 (2022).
29. Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I. & Heyn, H. SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Res.* **49**, e50 (2021).
30. Yan, G. & Li, J. J. scReadSim: a single-cell multi-omics read simulator. Preprint at *bioRxiv* https://doi.org/10.1101/2022.05.29.493924 (2022).
31. Cao, K., Hong, Y. & Wan, L. Manifold alignment for heterogeneous single-cell multi-omics data integration using Pamona. *Bioinformatics* **38**, 211–219 (2022).
32. Argelaguet, R., Cuomo, A. S. E., Stegle, O. & Marioni, J. C. Computational principles and challenges in single-cell data integration. *Nat. Biotechnol.* **39**, 1202–1215 (2021).
33. Fang, J. et al. Clustering deviation index (CDI): a robust and accurate internal measure for evaluating scRNA-seq data clustering. *Genome Biol.* **23**, 269 (2022).
34. Duò, A., Robinson, M. D. & Soneson, C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Res.* **7**, 1441 (2018).
35. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
36. Ji, Z. & Ji, H. TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.* **44**, e117 (2016).

https://doi.org/10.1038/s41587-023-01772-1

Methods

The generative model of scDesign3

Mathematical notations of scDesign3's training data. The training data of scDesign3 contain three matrices: a cell-by-feature matrix (for example, features are genes or chromatin regions), a cell-by-state-covariate matrix (for example, cell-state covariates include the cell type, pseudotime or spatial coordinate) and an optional cell-by-design-covariate matrix (for example, design covariates include the batch or condition).

Mathematically, first, we denote by $\mathbf{Y} = [Y_{ij}] \in \mathbb{R}^{n\times m}$ the cell-by-feature matrix with n cells as rows, m features as columns and Y_{ij} as the measurement of feature j in cell i. For single-cell sequencing data, $\mathbf{Y}$ is often a count matrix (that is, $\mathbf{Y} \in \mathbb{N}^{n\times m}$, with Y_{ij} indicating the read or unique molecular identifier (UMI) count of feature j in cell i); then the sequencing depth (that is, the total number of reads or UMIs) is $N = \sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ij}$.

Second, we denote by $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n]^\top \in \mathbb{R}^{n\times p}$ the cell-by-state-covariate matrix with n cells as rows and p cell-state covariates as columns. In $\mathbf{X}$, the ith row $\mathbf{x}_i \in \mathbb{R}^p$ is cell i's state covariate vector. Typical cell-state covariates include the cell type ($p = 1$ categorical variable), the cell pseudotime in p lineage trajectories (p continuous variables) and the two- or three-dimensional cell spatial locations ($p = 2$ or 3 continuous variables).

Third, we denote by $\mathbf{Z} = [\mathbf{z}_1, \cdots, \mathbf{z}_n]^\top \in \mathbb{R}^{n\times q}$ the cell-by-design-covariate matrix with n cells as rows and q design covariates as columns. In $\mathbf{Z}$, the ith row $\mathbf{z}_i \in \mathbb{R}^q$ is cell i's design covariate vector. Example design covariates are categorical variables such as the batch and condition. Note that $\mathbf{Z}$ is optional: it is not required if cells are from a single condition and measured in a single batch. To simplify the discussion, in the following text, we write $\mathbf{Z} = [\mathbf{b}, \mathbf{c}]$, where $\mathbf{b} = (b_1, \ldots, b_n)^\top$ has $b_i \in \{1, \ldots, B\}$ representing cell i's batch, and $\mathbf{c} = (c_1, \ldots, c_n)^\top$ has $c_i \in \{1, \ldots, C\}$ representing cell i's condition.

Modeling features' marginal distributions. For each feature $j = 1, \ldots, m$ in every cell $i = 1, \ldots, n$, the measurement Y_{ij}—conditional on cell i's state covariates $\mathbf{x}_i$ and design covariates $\mathbf{z}_i = (b_i, c_i)^\top$—is assumed to follow a distribution $F_j(\cdot|\mathbf{x}_i, \mathbf{z}_i; \mu_{ij}, \sigma_{ij}, p_{ij})$, which is specified as the generalized additive model for location, scale and shape (GAMLSS)[37] (that is, the distribution family F_j depends on feature j only, but the parameters μ_{ij}, σ_{ij} and p_{ij} depend on both feature j and cell i):

$$\begin{cases} Y_{ij}|\mathbf{x}_i, \mathbf{z}_i & \overset{\text{ind}}{\sim} F_j(\cdot|\mathbf{x}_i, \mathbf{z}_i; \mu_{ij}, \sigma_{ij}, p_{ij}) \\ \theta_j(\mu_{ij}) & = \alpha_{j0} + \alpha_{jb_i} + \alpha_{jc_i} + f_{jc_i}(\mathbf{x}_i) \\ \log(\sigma_{ij}) & = \beta_{j0} + \beta_{jb_i} + \beta_{jc_i} + g_{jc_i}(\mathbf{x}_i) \\ \text{logit}(p_{ij}) & = \gamma_{j0} + \gamma_{jb_i} + \gamma_{jc_i} + h_{jc_i}(\mathbf{x}_i) \end{cases}, \quad (1)$$

where $\theta_j(\cdot)$ denotes feature j's specific link function of the mean parameter μ_{ij}, depending on F_j (Supplementary Table 3); σ_{ij} denotes the scale parameter (for example, standard deviation or dispersion); and p_{ij} denotes the zero-inflation proportion parameter. Note that μ_{ij}, σ_{ij} and p_{ij} do not always coexist, depending on the form of F_j (Supplementary Table 3). To ensure model identifiability, for $j = 1, \ldots, m$, we set $\alpha_{jb_i} = \beta_{jb_i} = \gamma_{jb_i} = 0$ when $b_i = 1$ and $\alpha_{jc_i} = \beta_{jc_i} = \gamma_{jc_i} = 0$ when $c_i = 1$.

$\theta_j(\mu_{ij})$ is assumed to have feature j's specific intercept α_{j0}, batch b_i's effect α_{jb_i} (specific to feature j), condition c_i's effect α_{jc_i} (specific to feature j) and cell-state covariates $\mathbf{x}_i$'s effect $f_{jc_i}(\mathbf{x}_i)$ (specific to feature j and condition c_i).

$\log(\sigma_{ij})$ is assumed to have feature j's specific intercept β_{j0}, batch b_i's effect β_{jb_i} (specific to feature j), condition c_i's effect β_{jc_i} (specific to feature j) and cell-state covariates $\mathbf{x}_i$'s effect $g_{jc_i}(\mathbf{x}_i)$ (specific to feature j and condition c_i).

$\text{logit}(p_{ij})$ is assumed to have feature j's specific intercept γ_{j0}, batch b_i's effect γ_{jb_i} (specific to feature j), condition c_i's effect γ_{jc_i} (specific to feature j) and cell-state covariates $\mathbf{x}_i$'s effect $h_{jc_i}(\mathbf{x}_i)$ (specific to feature j and condition c_i).

For $\theta_j(\mu_{ij})$, $\log(\sigma_{ij})$ and $\text{logit}(p_{ij})$, the interaction effects are considered between the condition and cell-state covariates, but not between the batch and cell-state covariates. This modeling choice is made based on empirical observations and the simplicity preference[38].

Note that if only the mean parameter μ_{ij} is assumed to depend on the state covariates $\mathbf{x}_i$, batch b_i and condition c_i, then the GAMLSS degenerates to a GAM[39].

Depending on the modality of feature j (for example, a gene's UMI count), scDesign3 specifies F_j to be one of the six distributions: Gaussian (Normal), Bernoulli, Poisson, Negative Binomial (NB), Zero-inflated Poisson (ZIP) and Zero-inflated Negative Binomial (ZINB); see Supplementary Table 3 for the specifications. Different specifications of F_j correspond to different link functions $\theta_j(\cdot)$ and parameters; see Supplementary Table 3 for the details.

Depending on cell i's cell-state covariates $\mathbf{x}_i$, scDesign3 specifies the functions $f_{jc_i}(\cdot), g_{jc_i}(\cdot)$ and $h_{jc_i}(\cdot)$ in the corresponding forms. See Supplementary Table 4 for the details. Below are the three typical forms of $f_{jc_i}(\cdot)$.

(1) When the cell-state covariate is the cell type (out of a total of K_C cell types) and $\mathbf{X} = (x_1, \ldots, x_n)^\top$ is a 1-column matrix with $x_i \in \{1, \ldots, K_C\}$

$$f_{jc_i}(x_i) = \alpha_{jc_i x_i},$$

which corresponds to cell-type x_i's effect on feature j in condition c_i. Note that for identifiability, $\alpha_{jc_i x_i} = 0$ if $c_i = 1$ or $x_i = 1$.

(2) When the cell-state covariates are the cell pseudotimes in p lineage trajectories, that is, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ with x_{il} indicating cell i's pseudotime in the lth lineage trajectory

$$f_{jc_i}(\mathbf{x}_i) = \sum_{l=1}^{p}\sum_{k=1}^{K} b_{jc_i lk}(x_{il})\beta_{jc_i lk},$$

where $\sum_{k=1}^{K} b_{jc_i lk}(\cdot)\beta_{jc_i lk}$ is a cubic spline function for pseudotime in the lth lineage. This formulation means that feature j under condition c_i has a specific smooth pattern in lineage l. The exact choice K, the dimension of the basis governing the flexibility of f_{jc_i}, is not critical as long as K is not too small (because automatic penalization would be used in the estimation of f_{jc_i} by the R package `mgcv`, which is used in the R package `gamlss`[39]); we set $K = 10$ as default; K cannot be larger than the number of data points.

(3) When the cell-state covariates are two-dimensional spatial locations, that is, $\mathbf{x}_i = (x_{i1}, x_{i2})^\top$ indicating cell i's two-dimensional spatial coordinates

$$f_{jc_i}(\mathbf{x}_i) = f^{\text{GP}}_{jc_i}(x_{i1}, x_{i2}, K),$$

a low-rank Gaussian process smoother described in refs. 39,40, where K is the dimension of the basis governing the flexibility of f_{jc_i}. This formulation means that feature j under condition c_i has a smooth two-dimensional function (that is, a surface). The exact choice K is not critical as long as K is large (because automatic penalization would be used in the estimation of f_{jc_i} by the R package `mgcv`, which is used in the R package `gamlss`[39]); we set $K = 400$ as default; K cannot be larger than the number of data points.

The distribution of $(Y_{ij}|\mathbf{x}_i, \mathbf{z}_i)$ in equation (1) is fitted by the function `gamlss()` in the R package `gamlss` (v.5.4-3) or the function `gam()` in the R package `mgcv` (v.1.8-40). The fitted distribution is denoted as $\hat{F}_j(\cdot|\mathbf{x}_i, \mathbf{z}_i), i = 1, \ldots, n; j = 1, \ldots, m$.

Modeling features' joint distribution. For cell $i = 1, \ldots, n$, we denote its measurements of the m features as a random vector $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im})^\top$,

https://doi.org/10.1038/s41587-023-01772-1

whose joint distribution—conditional on cell i's state covariates $\mathbf{x}_i$ and design covariates $\mathbf{z}_i$—is denoted as $F(\cdot|\mathbf{x}_i, \mathbf{z}_i) : \mathbb{R}^m \to [0,1]$. The section 'Modeling features' marginal distributions' specifies $F_j(\cdot|\mathbf{x}_i, \mathbf{z}_i)$, the distribution of $(Y_{ij}|\mathbf{x}_i, \mathbf{z}_i)$, $j = 1, \ldots, m$. In scDesign3, the joint cumulative distribution function (CDF) $F(\cdot|\mathbf{x}_i, \mathbf{z}_i)$ is modeled from the marginal CDFs $F_1(\cdot|\mathbf{x}_i, \mathbf{z}_i), \ldots, F_m(\cdot|\mathbf{x}_i, \mathbf{z}_i)$ using the copula $C(\cdot|\mathbf{x}_i, \mathbf{z}_i)$: $[0,1]^m \to [0,1]$:

$$F(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i) = C\left(F_1(y_{i1}|\mathbf{x}_i, \mathbf{z}_i), \cdots, F_m(y_{im}|\mathbf{x}_i, \mathbf{z}_i)|\mathbf{x}_i, \mathbf{z}_i\right),$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{im})^\top$ is a realization of $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im})^\top$.

The copula $C(\cdot|\mathbf{x}_i, \mathbf{z}_i)$ can be (1) the Gaussian copula or (2) the vine copula, specified below.

The Gaussian copula is defined as:

$$\begin{aligned} &C\left(F_1(y_{i1}|\mathbf{x}_i, \mathbf{z}_i), \cdots, F_m(y_{im}|\mathbf{x}_i, \mathbf{z}_i)|\mathbf{x}_i, \mathbf{z}_i\right) \\ &= \Phi_m\left(\Phi^{-1}(F_1(y_{i1}|\mathbf{x}_i, \mathbf{z}_i)), \cdots, \Phi^{-1}(F_m(y_{i1}|\mathbf{x}_i, \mathbf{z}_i)); \mathbf{R}(\mathbf{x}_i, \mathbf{z}_i)\right), \end{aligned}$$

where Φ^{-1} denotes the inverse of the CDF of the standard Gaussian distribution, and $\Phi_m(\cdot\,; \mathbf{R}(\mathbf{x}_i, \mathbf{z}_i))$ denotes the CDF of an m-dimensional Gaussian distribution with a zero mean vector and a covariance matrix equal to the correlation matrix $\mathbf{R}(\mathbf{x}_i, \mathbf{z}_i)$.

An issue with the Gaussian copula is that the likelihood calculation is not straightforward in the high-dimensional case when m is large and the sample correlation matrix $\hat{\mathbf{R}}(\mathbf{x}_i, \mathbf{z}_i)$, as an estimator of $\mathbf{R}(\mathbf{x}_i, \mathbf{z}_i)$, is not invertible. Then, the likelihood cannot be computed based on $\hat{\mathbf{R}}(\mathbf{x}_i, \mathbf{z}_i)$. To address this issue, we consider the vine copula.

The vine copula is a way to 'decompose' a high-dimensional copula into a sequence of bivariate copulas, in which every pair of features is modeled as a bivariate Gaussian distribution. In short, the vine copula provides a regular vine (R-vine) structure that uses conditioning to sequentially decompose an m-dimensional copula into a sequence of bivariate copulas; then the m-dimensional copula density function is approximated by the product of the bivariate copula density functions[41]. The vine copula is advantageous to the Gaussian copula because it enables the likelihood calculation in the high-dimensional case. A detailed definition of the vine copula is in Supplementary Methods.

To estimate $C(\cdot|\mathbf{x}_i, \mathbf{z}_i)$ as either the Gaussian or vine copula, we use the plug-in approach that takes the estimated $\hat{F}_1(\cdot|\mathbf{x}_i, \mathbf{z}_i), \ldots, \hat{F}_m(\cdot|\mathbf{x}_i, \mathbf{z}_i)$ from the section 'Modeling features' marginal distributions'. Specifically, when $\hat{F}_j(\cdot|\mathbf{x}_i, \mathbf{z}_i)$ is a continuous distribution, each observed y_{ij} is transformed as $u_{ij} = \hat{F}_j(y_{ij}|\mathbf{x}_i, \mathbf{z}_i)$. When $\hat{F}_j(\cdot|\mathbf{x}_i, \mathbf{z}_i)$ is a discrete distribution with the support on non-negative integers (for example, the Poisson distribution), $u_{1j}, \ldots, u_{nj}$ follow a discrete distribution. Since the Gaussian and vine copulas assume that features follow continuous distributions, we use the distributional transformation as in ref. 16:

$$u_{ij} = (1 - v_{ij})\hat{F}_j(y_{ij} - 1|\mathbf{x}_i, \mathbf{z}_i) + v_{ij}\hat{F}_j(y_{ij}|\mathbf{x}_i, \mathbf{z}_i), \quad y_{ij} = 1, 2, \ldots,$$

where v_{ij}'s are sampled independently from Uniform[0, 1], $i = 1, \ldots, n$; $j = 1, \ldots, m$. To unify and simplify our notations, we write $u_{ij} = \tilde{F}_j(y_{ij}|\mathbf{x}_i, \mathbf{z}_i)$, where $\tilde{F}_j(\cdot|\mathbf{x}_i, \mathbf{z}_i)$ is the CDF of a continuous distribution.

Then, $C(\cdot|\mathbf{x}_i, \mathbf{z}_i)$ is estimated from $\mathbf{u}_1, \ldots, \mathbf{u}_n$, where $\mathbf{u}_i = (u_{i1}, \ldots, u_{im})^\top$. For the Gaussian copula, we use the function `cora()` in the R package `Rfast` (v.2.0.6); specifically, $\hat{\mathbf{R}}(\mathbf{x}_i, \mathbf{z}_i)$ is the sample correlation matrix of $\{\Phi^{-1}(\mathbf{u}_j)$: $(\mathbf{x}_j, \mathbf{z}_j)$ is in a predefined-sized neighborhood of $(\mathbf{x}_i, \mathbf{z}_i)\}$, where $\Phi^{-1}(\mathbf{u}_i) = (\Phi^{-1}(u_{i1}), \ldots, \Phi^{-1}(u_{im}))^\top$. For the vine copula, we use the function `vinecop()` in R package `rvinecoplib` (v.0.6.2.1.1).

Then, the estimated joint distribution $\hat{F}(\cdot|\mathbf{x}_i, \mathbf{z}_i)$ is

$$\hat{F}(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i) = \hat{C}\left(\tilde{F}_1(y_{i1}|\mathbf{x}_i, \mathbf{z}_i), \cdots, \tilde{F}_m(y_{im}|\mathbf{x}_i, \mathbf{z}_i)|\mathbf{x}_i, \mathbf{z}_i\right). \tag{2}$$

Model likelihood, AIC and BIC. Given equation (2), the estimated probability density function of cell i's m-dimensional feature vector $\mathbf{y}_i$, conditional on the cell-state covariates $\mathbf{x}_i$ and the design covariates $\mathbf{z}_i$, is

$$\hat{f}(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i) = \hat{c}\left(\tilde{F}_1(y_{i1}|\mathbf{x}_i, \mathbf{z}_i), \cdots, \tilde{F}_m(y_{im}|\mathbf{x}_i, \mathbf{z}_i)|\mathbf{x}_i, \mathbf{z}_i\right)\prod_{j=1}^{m}\tilde{f}_j(y_{ij}|\mathbf{x}_i, \mathbf{z}_i),$$

where $\hat{c}(\cdot|\mathbf{x}_i, \mathbf{z}_i)$ is the probability density function of $\hat{C}(\cdot|\mathbf{x}_i, \mathbf{z}_i)$, and $\tilde{f}_j(\cdot|\mathbf{x}_i, \mathbf{z}_i)$ is the probability density function of $\tilde{F}_j(\cdot|\mathbf{x}_i, \mathbf{z}_i)$. Hence, the log-likelihood is

$$\begin{aligned} \ell &= \sum_{i=1}^{n}\log\hat{f}(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i) \\ &= \sum_{i=1}^{n}\log\hat{c}\left(\tilde{F}_1(y_{i1}|\mathbf{x}_i, \mathbf{z}_i), \cdots, \tilde{F}_m(y_{im}|\mathbf{x}_i, \mathbf{z}_i)|\mathbf{x}_i, \mathbf{z}_i\right) + \sum_{i=1}^{n}\sum_{j=1}^{m}\log\tilde{f}_j(y_{ij}|\mathbf{x}_i, \mathbf{z}_i) \\ &= \ell^{\text{Copula}} + \ell^{\text{Marginal}}, \end{aligned}$$

so the log-likelihood ℓ can be written as the sum of a copula log-likelihood

$$\ell^{\text{Copula}} = \sum_{i=1}^{n}\log\hat{c}\left(\tilde{F}_1(y_{i1}|\mathbf{x}_i, \mathbf{z}_i), \cdots, \tilde{F}_m(y_{im}|\mathbf{x}_i, \mathbf{z}_i)|\mathbf{x}_i, \mathbf{z}_i\right)$$

and a marginal log-likelihood

$$\ell^{\text{Marginal}} = \sum_{i=1}^{n}\sum_{j=1}^{m}\log\tilde{f}_j(y_{ij}|\mathbf{x}_i, \mathbf{z}_i).$$

Given k model parameters and n cells (that is, the sample size n is the number of cells), the AIC and BIC are

$$\begin{aligned} \text{AIC} &= 2k - 2\ell; \\ \text{BIC} &= 2k\log(n) - 2\ell, \end{aligned}$$

so smaller AIC and BIC values indicate better goodness-of-fit of a model to data.

Because of the likelihood decomposition, the AIC and BIC are also decomposable

$$\begin{aligned} \text{AIC} &= \text{AIC}^{\text{Copula}} + \text{AIC}^{\text{Marginal}}; \\ \text{BIC} &= \text{BIC}^{\text{Copula}} + \text{BIC}^{\text{Marginal}}, \end{aligned}$$

where $\text{AIC}^{\text{Copula}}$ and $\text{BIC}^{\text{Copula}}$ only include the number of parameters in $\hat{c}(\cdot|\mathbf{x}_i, \mathbf{z}_i)$, and $\text{AIC}^{\text{Marginal}}$ and $\text{BIC}^{\text{Marginal}}$ only include the total number of parameters in $\tilde{f}_1(\cdot|\mathbf{x}_i, \mathbf{z}_i), \ldots, \tilde{f}_m(\cdot|\mathbf{x}_i, \mathbf{z}_i)$.

Synthetic data generation by scDesign3

To generate a synthetic cell-by-feature matrix $\mathbf{Y}' \in \mathbb{R}^{n' \times m}$, which contains n' synthetic cells and the same m features as in the training data, scDesign3 allows the specification of a cell-by-state-covariate matrix $\mathbf{X}' \in \mathbb{R}^{n' \times p}$ and an optional cell-by-design-covariate matrix $\mathbf{Z}' \in \mathbb{N}^{n' \times q}$ (depending on whether the training data have $\mathbf{Z}$) for the n' synthetic cells. Note that $\mathbf{X}'$ and $\mathbf{Z}'$ can be specified by users, generated by resampling the rows of $\mathbf{X}$ and $\mathbf{Z}$, or sampled from some generative models of the rows of $\mathbf{X}$ and $\mathbf{Z}$.

Given $\mathbf{X}$, $\mathbf{Z}$ and the fitted distributions in sections 'Modeling features' marginal distributions' and 'Modeling features' joint distribution', scDesign3 samples n' synthetic cells in the following steps.

First, for each synthetic cell i', given its cell-state covariates $\mathbf{x}_{i'}$ and design covariates $\mathbf{z}_{i'}$, we independently sample an m-dimensional vector (with values in [0, 1]) from the m-dimensional copula estimated in the section 'Modeling features' joint distribution':

$$(U_{i'1}, \ldots, U_{i'm})^\top \sim \hat{C}(\cdot|\mathbf{x}_{i'}, \mathbf{z}_{i'}), \quad i' = 1, \ldots, n'.$$

Second, based on the m features' fitted marginal distributions in the section 'Modeling features' marginal distributions', we calculate the conditional distribution of $Y_{i'j}$, the measurement of feature j in synthetic cell i', given the synthetic cell's cell-state covariates $\mathbf{x}_{i'}$ and design covariates $\mathbf{z}_{i'} = (b_{i'}, c_{i'})^\top$, where $b_{i'} \in \{1, \ldots, B\}$ and $c_{i'} \in \{1, \ldots, C\}$; that is, the distribution of $Y_{i'j}|\mathbf{x}_{i'}, \mathbf{z}_{i'}$:

$$\hat{F}_j(\cdot|\mathbf{x}_{i'}, \mathbf{z}_{i'}) = F_j(\cdot|\mathbf{x}_{i'}, \mathbf{z}_{i'}; \hat{\mu}_{i'j}, \hat{\sigma}_{i'j}, \hat{p}_{i'j}),$$

where

$$\begin{cases} \theta(\hat{\mu}_{i'j}) & = \hat{\alpha}_{j0} + \hat{\alpha}_{jb_{i'}} + \hat{\alpha}_{jc_{i'}} + \hat{f}_{jc_{i'}}(\mathbf{x}_{i'}), \\ \log(\hat{\sigma}_{i'j}) & = \hat{\beta}_{j0} + \hat{\beta}_{jb_{i'}} + \hat{\beta}_{jc_{i'}} + \hat{g}_{jc_{i'}}(\mathbf{x}_{i'}), \\ \text{logit}(\hat{p}_{i'j}) & = \hat{\gamma}_{j0} + \hat{\gamma}_{jb_{i'}} + \hat{\gamma}_{jc_{i'}} + \hat{h}_{jc_{i'}}(\mathbf{x}_{i'}). \end{cases}$$

Note that $\hat{\mu}_{i'j}, \hat{\sigma}_{i'j}$ and $\hat{p}_{i'j}$ may not all be required, depending on the form of F_j (Supplementary Table 3).

Then, the m-dimensional feature vector of synthetic cell i' is $(Y_{i'1}, \ldots, Y_{i'm})^\top$, where

$$Y_{i'j} = \hat{F}_j^{-1}(U_{i'j}|\mathbf{x}_{i'}, \mathbf{z}_{i'}), \quad j = 1, \ldots, m.$$

Thanks to the parametric form of $\hat{F}_j(\cdot|\mathbf{x}_{i'}, \mathbf{z}_{i'})$, users can generate the synthetic data in their demand by modifying the parameters. For instance, if users want the expected sequencing depth of $\mathbf{Y}'$ to change from N (the sequencing depth of $\mathbf{Y}$) to N', they can scale the mean parameter; that is, the distribution of $Y_{i'j}|\mathbf{x}_{i'}, \mathbf{z}_{i'}$ becomes:

$$\hat{F}_j(\cdot|\mathbf{x}_{i'}, \mathbf{z}_{i'}) = F_j\left(\cdot \,\middle|\, \mathbf{x}_{i'}, \mathbf{z}_{i'}; \frac{N'}{N}\hat{\mu}_{i'j}, \hat{\sigma}_{i'j}, \hat{p}_{i'j}\right).$$

If users want to remove the batch effects, they can set

$$\hat{\alpha}_{jb_{i'}} = \hat{\beta}_{jb_{i'}} = \hat{\gamma}_{jb_{i'}} = 0,$$

for all $i' = 1, \ldots, n'; j = 1, \ldots, m$.

If users want to remove the condition effects, they can set

$$\begin{aligned} \hat{\alpha}_{jc_{i'}} &= \hat{\beta}_{jc_{i'}} = \hat{\gamma}_{jc_{i'}} = 0; \\ \hat{f}_{jc_{i'}}(\cdot) &= \hat{f}_{j1}(\cdot); \\ \hat{g}_{jc_{i'}}(\cdot) &= \hat{g}_{j1}(\cdot); \\ \hat{h}_{jc_{i'}}(\cdot) &= \hat{h}_{j1}(\cdot), \end{aligned}$$

for all $i' = 1, \ldots, n'; j = 1, \ldots, m$.

The comparison of scDesign, scDesign2 and scDesign3. Supplementary Table 1 lists a detailed comparison of scDesign3 with the previous two versions scDesign[24] and scDesign2[16]. Note that scDesign2 is a special case of scDesign3 for generating scRNA-seq data from discrete cell types.

Data analysis

Data preprocessing. Supplementary Table 2 lists the real datasets from 17 published studies, which were used in this study. Since scDesign3 can directly model count data, we did not perform data transformation (for example, logarithmic transformation) on the cell-by-feature count matrices.

For each cell-by-feature count matrix $\mathbf{Y}$ (except for the SCGEM-METH and SCGEM-RNA datasets), feature screening was used to select informative features and save computation time.

- For every scRNA-seq dataset (BATCH, EMBRYO, IFNB, MARROW, PANCREAS and the RNA data in CITE), we used the R package `scran` (v.1.20.1)[42] to select the top 1,000 highly variable genes.
- For the 10x scATAC-seq dataset (ATAC), we used the R package `Signac` (v.1.7.0)[43] to first obtain a cell-by-peak matrix and then select 1,133 differentially accessible peaks.
- For the sci-ATAC-seq (SCIATAC) dataset, the preprocessing and feature selection steps were described[30].
- For the 10x Visium datasets (ACINAR, OVARIAN and VISIUM), we used the R package `Seurat` (v.4.1.1)[44] to select the top 1,000 spatially variable genes.
- For the Slide-seq dataset (SLIDE), we selected the top 1,000 genes with the smallest P values outputted by `SPARK-X`[45].
- For the pair of single-cell and spatial datasets (MOB-SC and MOB-SP), we used the R package `scran` (v.1.20.1) to select the top 50 marker genes for each cell type in MOB-SC.
- For datasets MOUSE-CORTEX, MOUSE-VISUAL and ZHENGMIX4, we used the genes selected in the original studies[34,46].

For each dataset, the cell-by-state-covariate matrix $\mathbf{X}$ was from the original study (if the cell-state covariates are cell types or spatial locations) or inferred by the R package `Slingshot` (v.2.2.1)[35] (if the cell-state covariates are pseudotime values in trajectory lineages).

For each dataset, the optional cell-by-design-covariate matrix $\mathbf{Z}$ was from the original study if available.

Dimensionality reduction and visualization. To compare scDesign3's synthetic data with real test data, we used the R package `irlba` (v.2.3.5) for principal component analysis (PCA), that is, to calculate the top 50 principal components of the test cell-by-feature matrix (after log-transformation); next, we used the R package `UMAP` (v.0.2.8.0) to project the test cells from the 50-dimensional principal component space to the two-dimensional UMAP space. Then, we applied the same PCA-UMAP projection to scDesign3's synthetic cells using the R function `predict()`. Using the same projection ensures that the test cells and synthetic cells are embedded in the same two-dimensional space and thus comparable.

Unless otherwise noted, the figures were made by the R package `ggplot2` (v.3.3.6). The coverage plot in Fig. 1g was generated by `IGV` (v.2.12.3).

Evaluation metrics.

- mLISI: To measure the similarity between test cells and synthetic cells in the two-dimensional space, we used mLISI[25] as the metric. Specifically, if a cell's neighboring cells are from one group (for example, test cells or synthetic cells), the cell's local inverse Simpson's index (LISI) is 1; otherwise, if a cell's neighboring cells comprise two groups equally, the cell's LISI is 2. The mLISI is the average of all cells' LISIs. Hence, an mLISI close to 2 means that the test cells and synthetic cells are well mixed. The mLISI is calculated by the function `evalIntegration()` in the R package `CellMixS` (v.1.8.0)[47].
- Pearson correlation between spatial patterns: To measure the per-feature similarity between real data and synthetic data when the cell-state covariates are spatial locations, we compared supervised learners trained on real data and synthetic data separately. In detail, for every feature (for example, gene), we conducted the following analysis. First, treating the feature as the outcome, we trained a flexible learner, the generalized boosted regression model (GBM), separately on real data and synthetic data to predict the feature's values from the cell-state covariates, using the R package `caret` (v.6.0-93). Second, we measured the Pearson correlation r between the two GBMs' predicted feature values from the synthetic data's spatial locations (note that the cell-state covariates could be replaced by a random sample from the location space). An r close to 1 means that the two GBMs are similar; that is, the feature's 'relationship' with spatial locations is similar in the real data and the synthetic data. If all features

have r values close to 1, we concluded that the synthetic data resemble the real data.

- Summary statistics: In Extended Data Figs. 1, 2, 4, 8 and 9 and Supplementary Figs. 1–5, we compared the distributions of eight feature-level, cell-level, feature-pair-level and cell-pair-level summary statistics between real data and synthetic data. Note that in scRNA-seq and spatial transcriptomics data, every gene is a feature; in scATAC-seq and sci-ATAC-seq data, every peak is a feature. The eight summary statistics are:
  (1) Mean of log expression (feature-level statistic): a feature's mean of $\log(\text{count}+1)$ values across all cells.
  (2) Variance of log expression (feature-level statistic): a feature's variance of $\log(\text{count}+1)$ values across all cells.
  (3) Feature detection frequency (feature-level statistic): a feature's proportion of nonzero counts across all cells.
  (4) Feature–feature correlation (feature-pair-level statistic): the correlation between two features' $\log(\text{count}+1)$ values across all cells.
  (5) Cell library size on the log scale (cell-level statistic): a cell's log-transformed total read or UMI count (that is, log per-cell sequencing depth).
  (6) Cell–cell distance (cell-pair-level statistic): the Euclidean distance between two cells in the 50-dimensional principal component space (constructed from the cell-by-gene $\log(\text{count}+1)$ matrix).
  (7) Cell detection frequency (cell-level statistic): a cell's proportion of nonzero counts across all features.
  (8) Cell–cell correlation (cell-pair-level statistic): the correlation between two cells' $\log(\text{count}+1)$ values across all features.

Feature–feature correlations were calculated for the top 100 highly expressed features in each real dataset and the corresponding synthetic datasets. To measure the similarity between the real and synthetic correlation matrices, we calculated the Pearson correlation r across all 100^2 entries of the correlation matrices.

Boxplots and scatter plots. The boxplots (Fig. 2f) were plotted using the function `geom_boxplot()` in the R package `ggplot2` (v.3.6.6). In each boxplot, the center horizontal line represents the median, the box limits represent the upper and lower quartiles, the whiskers cover the 1.5 × interquartile range and points are outliers. The P value was calculated by the two-sided Wilcoxon rank-sum test.

The scatter plots (Fig. 2e and Extended Data Fig. 10) were plotted using the function `geom_scatter()` in the R package `ggplot2` (v.3.6.6). In each scatter plot, the P value associated with the Spearman's correlation coefficient ρ was calculated by the one-sided test in the function `cor.test()` in the R package `stats` (v.4.4.2).

scDesign3's simulation of spot-resolution transcriptomics data with true cell-type proportions. To generate the synthetic spot-resolution spatial transcriptomics data with true cell-type proportions at each spot, we used a pair of an scRNA-seq dataset (MOB-SC) and a spatial transcriptomics dataset (MOB-SP) that measured the same biological sample (mouse olfactory bulb). The simulation procedure consists of three steps: the first two steps for parameter estimation from real data and the last step for data simulation.

First, we used scDesign3 to estimate each gene's mean expression level of each cell type (from scRNA-seq data) and the same gene's mean expression level at each spatial spot (from spatial transcriptomics data; Extended Data Fig. 7a, Step 1).

Second, using the four cell types' gene mean expression vectors (one vector per cell type; the cell types are the columns in Extended Data Fig. 7b; each vector's elements correspond to genes' mean expression levels in the cell type) as the reference data and the spatial spots' gene mean expression vectors (one vector per spot) as the query data, we used the cell-type decomposition method CIBERSORT[48,49] to estimate each spot's cell-type proportions (Fig. 1f, left, and Extended Data Fig. 7b, top row), which we then considered as the spot's true cell-type proportions in scDesign3's simulation. As a sanity check, we show CIBERSORT's fitted gene expression levels at each spot in Extended Data Fig. 7a, Step 2. Note that CIBERSORT could be replaced by other decomposition methods.

Third, we used scDesign3 to generate synthetic scRNA-seq data of the four cell types after training scDesign3 on the real scRNA-seq data. Then, we simulated spot-resolution transcriptomics data as follows. For each real spot, we sampled 100 cells from the four cell types based on the spot's true cell-type proportions. Specifically, if the true cell-type proportions are $p_1, \ldots, p_4$, then the numbers of cells sampled from the four cell types would be drawn from a multinomial distribution, Multinomial(100, $(p_1, \ldots, p_4)$). Then, we added the sampled cells' gene expression vectors and divided the summed vector by 10 to form the spot's gene expression vector (so every spot corresponds to 10 cells on average, consistent with real data) (Extended Data Fig. 7a, Step 3).

Using the synthetic spot-resolution spatial transcriptomics data, we benchmarked three spatial transcriptomics cell-type deconvolution algorithms: CARD[27], RCTD[28] and SPOTlight[29], using the R packages `CARD` (v.1.0), `spacexr` (v.2.1.6) and `SPOTlight` (v.1.0.1), respectively. We chose these three algorithms to demonstrate scDesign3's benchmarking functionality because of a published benchmark study[27], which found CARD and RCTD to have similarly good performance and to have outperformed SPOTlight. Hence, we considered CARD, RCTD and SPOTlight as representative algorithms to check if our benchmark results based on scDesign3 could be consistent with the published study that used an independent approach[27].

scDesign3's simulation of a multiomics dataset from single-omics datasets measuring different modalities. To simulate a multiomics dataset from real single-omics datasets with unmatched cells, scDesign3 relies on an integration method that projects single-omics data to a joint low-dimensional space. Then, scDesign3 considers each cell's low-dimensional embedding as the cell-state covariates in the modeling.

In Fig. 1j, we used an scRNA-seq dataset and a single-cell methylation dataset with unmatched cells. The two datasets' cells' joint low-dimensional embeddings were inferred by the integration method Pamona[31], which could be replaced by other integration methods. Then, we trained scDesign3 for each modality (RNA or methylation) using the low-dimensional embeddings of the modality's real cells. Finally, using the fitted models (one per modality), we generated a synthetic cell with both modalities from each real cell's low-dimensional embedding.

scDesign3's assessment of cell clusters' goodness-of-fit. To show that scDesign3 can assess the goodness-of-fit of cell clusters, we used the eight datasets from the R package `DuoClustering2018` (v.1.10.0), in which each dataset contains cell-type labels ('truth') and various clustering methods' results with varying numbers of clusters. The ARI, a 'supervised' measure calculated between each clustering result and cell-type labels, was used as the benchmark standard. Assuming the NB distribution in the scDesign3 model, we calculated scDesign3's marginal BIC (in the section 'Model likelihood, AIC and BIC'), an 'unsupervised' measure that uses only the clustering result but not the cell-type labels, for each clustering result in each dataset. We used scDesign3's marginal BIC because we observed that it better captured the goodness-of-fit of cell clusters than the scDesign3 BIC. A possible reason is that the scDesign3 BIC is dominated by the copula BIC, which largely reflects the number of parameters instead of the clustering goodness-of-fit.

In Extended Data Fig. 10b, we benchmarked scDesign3's marginal BIC against the ARI and found them to have negative correlations on the eight datasets consistently, suggesting that scDesign3's marginal

BIC is an effective assessment measure of clustering goodness-of-fit: a lower BIC indicates better goodness-of-fit.

scDesign3's assessment of cell pseudotimes' goodness-of-fit. To show that scDesign3 can assess the goodness-of-fit of cell pseudotimes, we used five synthetic datasets generated by the R package `dyngen` (v.1.0.3)[19] and three synthetic datasets generated by scDesign3; each dataset contains cells' true pseudotime values ('truth') ranging from 0 to 1. To generate perturbed pseudotimes with varying quality, we randomly replaced 0%, 10%, 20%, ..., 100% of truth pseudotime values with randomly sampled values from the Uniform[0, 1] distribution. We also considered the inferred pseudotimes by the R packages `Slingshot` (v.2.4.0), `Monocle3` (v.1.0.0) and `TSCAN` (v.1.34.0). The benchmark standard was the 'supervised' R^2 between the true pseudotime values and the perturbed or inferred pseudotime values. Using the NB distribution in the scDesign3 model, we calculated scDesign3's marginal BIC (in the section 'Model likelihood, AIC and BIC'), an 'unsupervised' measure that only uses the perturbed or inferred pseudotime values but not the true pseudotime values, for each set of perturbed or inferred pseudotime values in each dataset. We used scDesign3's marginal BIC because we observed that it better captured the goodness-of-fit of cell pseudotimes than the scDesign3 BIC. A possible reason is that the scDesign3 BIC is dominated by the copula BIC, which largely reflects the number of parameters instead of the pseudotime goodness-of-fit.

In Extended Data Fig. 10a, we benchmarked scDesign3's marginal BICs against the R^2 and found them to have negative correlations on the eight datasets consistently, suggesting that scDesign3's marginal BIC is an effective assessment measure of pseudotime goodness-of-fit: a lower BIC indicates better pseudotime goodness-of-fit.

scDesign3's assessment of inferred spatial locations' goodness-of-fit. To show that scDesign3 can assess the goodness-of-fit of inferred spatial locations, we used two single-cell resolution spatial transcriptomics datasets from Li et al.[46]. The two datasets contain all cells' measured spatial locations. Then, for each spatial transcriptomics dataset, we treated its cells' gene expression counts as a 'pseudo' scRNA-seq dataset, and we inputted this pseudo scRNA-seq data along with the original spatial transcriptomics dataset into Seurat (v.4.1.1), Tangram (v.1.0.0)[50] and novoSpaRc (v.0.4.3)[51]—as an integration task—to infer the spatial locations of the cells in the pseudo scRNA-seq dataset. This approach allowed us to evaluate the inferred spatial locations based on the true spatial locations in the original spatial transcriptomics dataset.

The inferred spatial locations by novoSpaRc contained a large proportion of overlapping locations and thus were not used in our assessment. For Seurat and Tangram, we used each method's inferred spatial locations along with the original gene expression counts to train scDesign3 (with the NB distribution; Supplementary Table 3) and calculate the likelihood, marginal AIC and marginal BIC (in the section 'Model likelihood, AIC and BIC'). Note that we only used the top 100 spatially variable genes defined by Moran's *I* statistic to train scDesign3 to save computational time. To evaluate the performance of scDesign3's unsupervised marginal AIC and BIC, we used the mean cosine similarity, a 'supervised' measure that averages all cells' absolute values of the cosine similarity (for each cell, the cosine similarity is calculated between the cell's true spatial location and inferred spatial location).

Additionally, for each dataset, we randomly shuffled 0%, 10%, 20%, ..., 100% of true spatial locations to obtain perturbed spatial locations with varying quality. Then, we calculated scDesign3's marginal AIC and BIC for the perturbed spatial locations.

In Extended Data Fig. 10c, we found that scDesign3's marginal AIC and the mean cosine similarity had negative correlations on the two datasets, suggesting that scDesign3's marginal AIC is an effective assessment measure of spatial locations' goodness-of-fit: a lower AIC indicates better goodness-of-fit. Note that AIC outperformed BIC in this case, possibly due to the reason that genes' spatial patterns are complex and thus need complex models.

Implementation of other simulators. We compared scDesign3 with multiple representative scRNA-seq simulators including scGAN, muscat, SPARSim and ZINB-WaVE.

- For scGAN, we used the docker and the tutorial available at https://github.com/imsb-uke/scGAN (access date: 7 February 2022) to simulate synthetic data.
- For muscat, we first used the R function `prepSim()` to process the training dataset. Then, we ran the R function `simData()` to simulate a synthetic dataset based on the processed training dataset and the cell-level information (such as cell types and experimental conditions) in the training dataset. Both functions are from the R package `muscat` (v.1.6.0).
- For SPARSim, we first used the `SPARSim_create_simulation_parameter()` function to obtain the parameters for each group of cells in the training dataset, whose cells were grouped by cell types, experimental conditions or batches. The three required input parameters for the function—`intensity`, `variability` and `library_size`—were obtained from the functions `SPARSim_estimate_intensity()`, `SPARSim_estimate_variability()` and `SPARSim_estimate_library_size()`, respectively, for each cell group. Then, we ran the `SPARSim_simulation()` function with the input parameters from the previous step to generate synthetic data. All functions are from the R package `SPARSim` (v.0.9.5).
- For ZINB-WaVE, we used the `zinbFit()` function from the R package `zinbwave` (v.1.15.3), with the count matrix and cell-type labels as inputs.

Reporting summary

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

Data availability

All datasets used in the study are publicly available. Supplementary Table 2 lists the datasets from 17 published studies (sources included). The preprocessed datasets are available in the Zenodo repository at https://doi.org/10.5281/zenodo.7110761[52].

Code availability

The scDesign3 package is available at https://github.com/SONGDONGYUAN1994/scDesign3. The comprehensive tutorials are available at https://songdongyuan1994.github.io/scDesign3/docs/index.html. In the tutorials, we described the input and output formats, model parameters and exemplary datasets for each functionality of scDesign3. The source code for reproducing the results is available in the Zenodo repository at https://doi.org/10.5281/zenodo.7110761[52].

References

37. Stasinopoulos, D. M. & Rigby, R. A. Generalized additive models for location scale and shape (GAMLSS) in R. *J. Stat. Softw.* **23**, 1–46 (2008).
38. Zhang, Y., Parmigiani, G. & Johnson, W. E. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom. Bioinform.* **2**, lqaa078 (2020).
39. Wood, S. N. *Generalized Additive Models: An Introduction with R* (Chapman and Hall/CRC, 2006).
40. Kammann, E. E. & Wand, M. P. Geoadditive models. *J. R. Stat. Soc. C* **52**, 1–18 (2003).
41. Czado, C. *Analyzing Dependent Data with Vine Copulas* (Springer, 2019).

42. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res.* **5**, 2122 (2016).
43. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).
44. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
45. Zhu, J., Sun, S. & Zhou, X. SPARK-X: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome Biol.* **22**, 184 (2021).
46. Li, B. et al. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nat. Methods* **19**, 662–670 (2022).
47. Lütge, A. et al. CellMixS: quantifying and visualizing batch effects in single-cell RNA-seq data. *Life Sci. Alliance* **4**, e202001004 (2021).
48. Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
49. Zeng, D. et al. IOBR: multi-omics immuno-oncology biological research to decode tumor microenvironment and signatures. *Front. Immunol.* **12**, 687975 (2021).
50. Biancalani, T. et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nat. Methods* **18**, 1352–1362 (2021).
51. Moriel, N. et al. Novosparc: flexible spatial reconstruction of single-cell gene expression with optimal transport. *Nat. Protoc.* **16**, 4177–4200 (2021).
52. Song, D., Wang, Q. & Li, J. J. scDesign3 generates realistic in silico data for multimodal single-cell and spatial omics. *Zenodo* https://doi.org/10.5281/zenodo.7110761 (2022).

Acknowledgements

We appreciate the comments and feedback from the members of the Junction of Statistics and Biology at UCLA (http://jsb.ucla.edu).

This work was supported by the following grants: National Science Foundation grants no. DBI-1846216 and no. DMS-2113754, NIH/NIGMS grants no. R01GM120507 and no. R35GM140888, Johnson & Johnson WiSTEM2D Award, the Sloan Research Fellowship, the UCLA David Geffen School of Medicine W. M. Keck Foundation Junior Faculty Award and the Chan-Zuckerberg Initiative Single-Cell Biology Data Insights Grant (to J.J.L.). J.J.L. was a fellow at the Radcliffe Institute for Advanced Study at Harvard University in 2022–2023 while she was writing this paper.

Author contributions

D.S. and J.J.L. conceived of the study. D.S., Q.W. and J.J.L. wrote the paper. D.S. and Q.W. developed the scDesign3 R package. D.S. and Q.W. performed data analysis with assistance from G.Y. and T.L. D.S. and T.S. discussed the scDesign3 method design at the beginning of the study.

Competing interests

The authors declare no competing interests.

Additional information

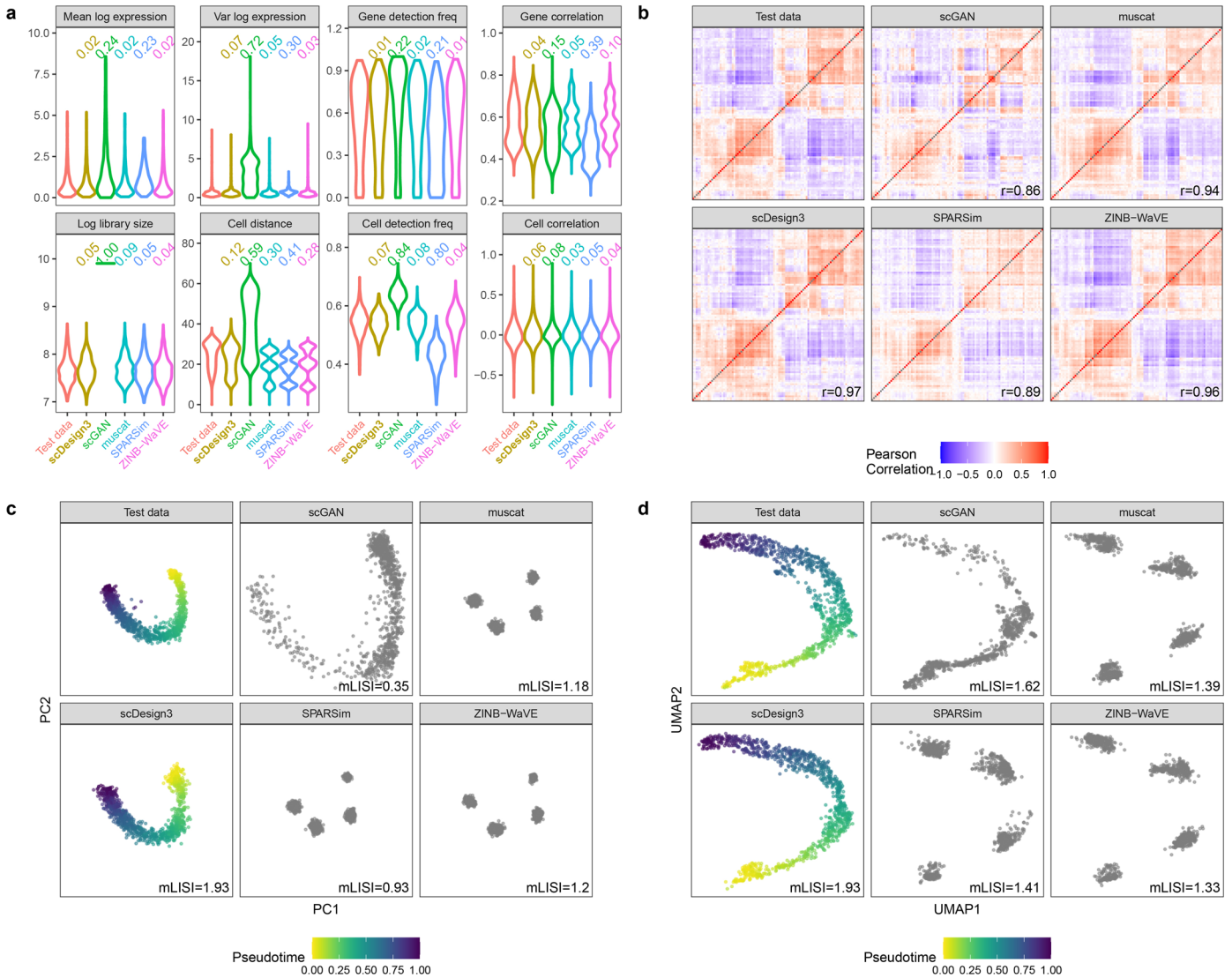
Extended data is available for this paper at https://doi.org/10.1038/s41587-023-01772-1.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-023-01772-1.

Correspondence and requests for materials should be addressed to Jingyi Jessica Li.

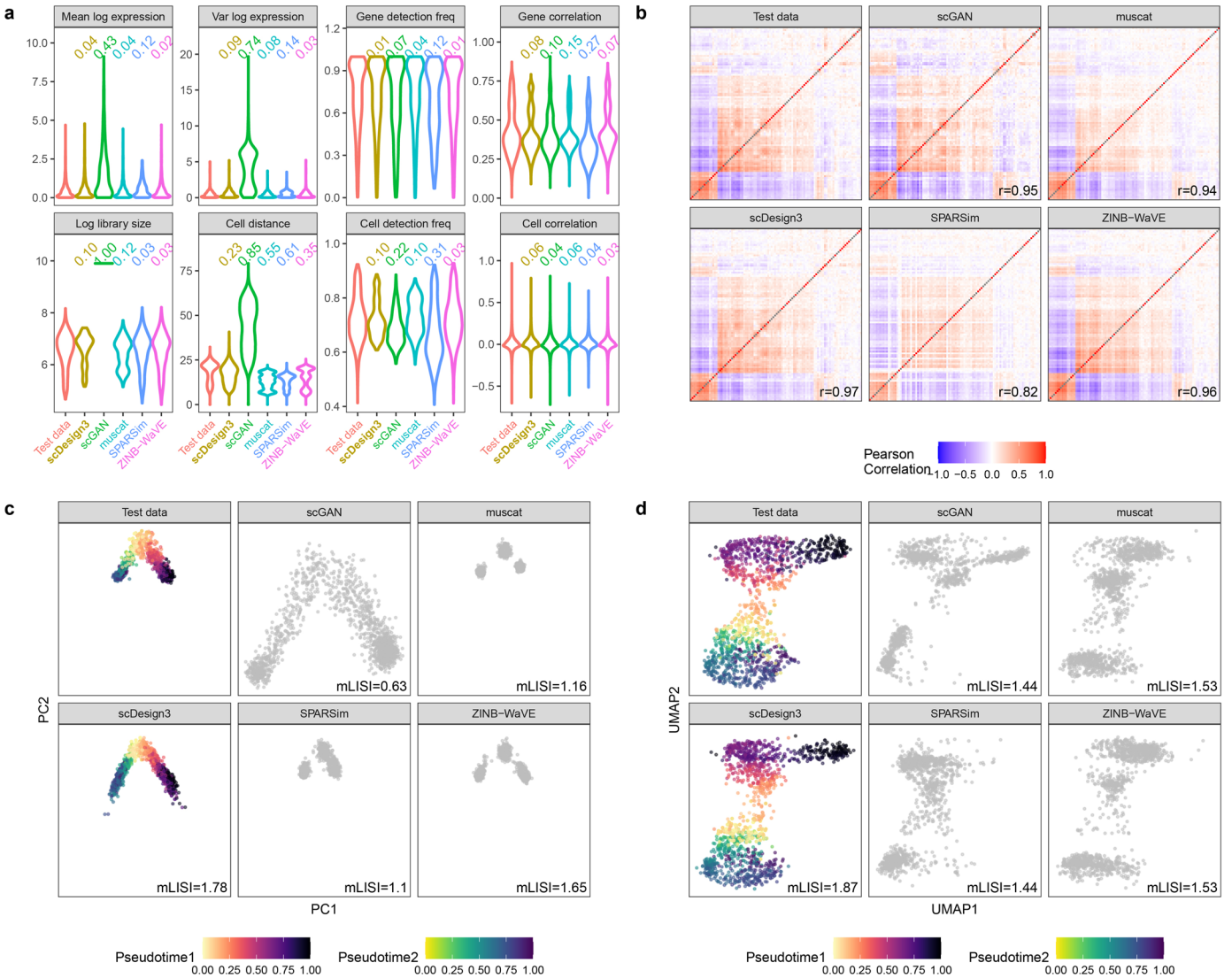
Peer review information *Nature Biotechnology* thanks Kin Fai Au and Jean Yang for their contribution to the peer review of this work.

https://doi.org/10.1038/s41587-023-01772-1



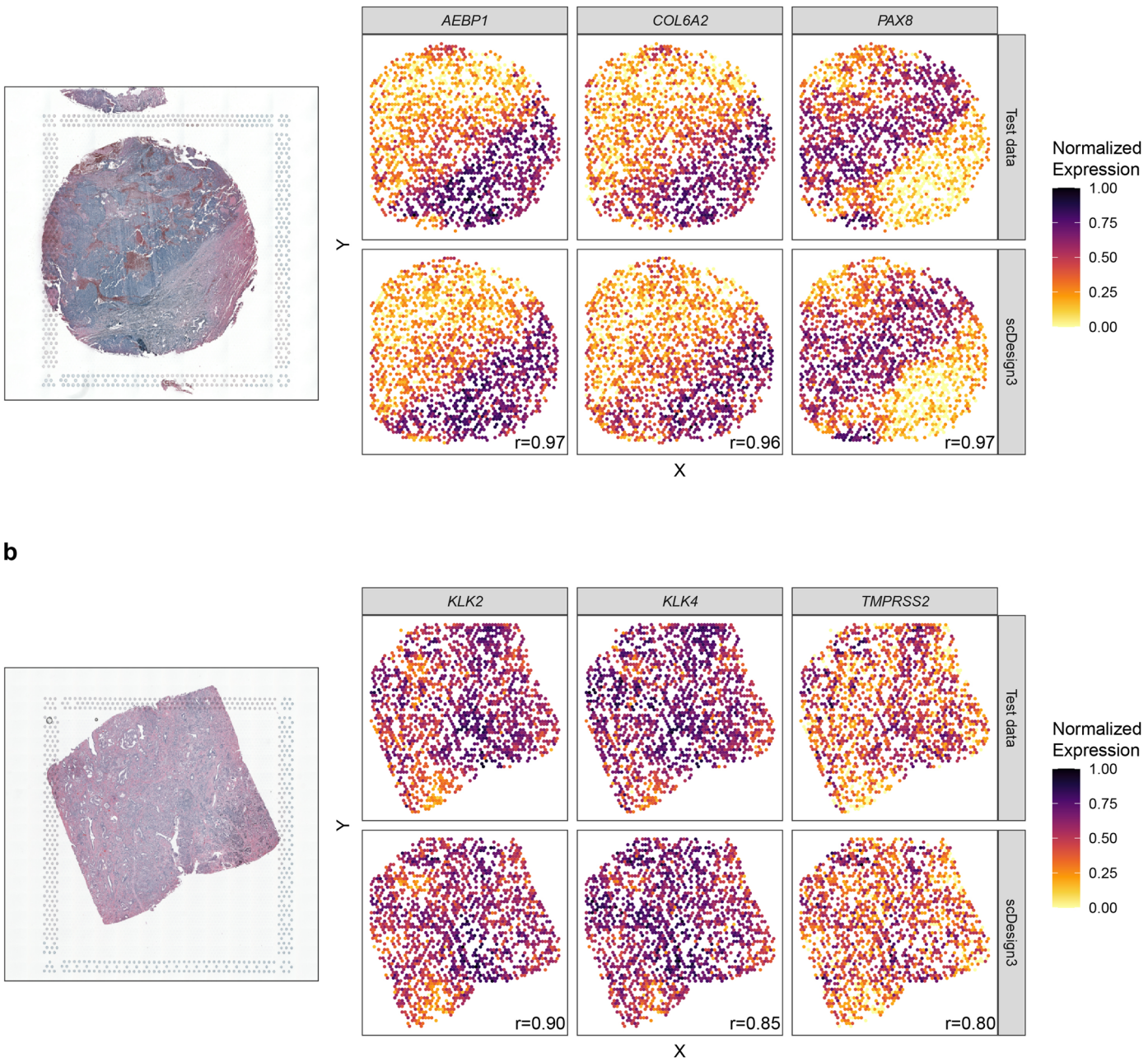
Extended Data Fig. 1 | Benchmarking scDesign3 against four existing scRNA-seq simulators (scGAN, muscat, SPARSim, and ZINB-WaVE) for generating scRNA-seq data from a single trajectory (mouse pancreatic endocrinogenesis; dataset PANCREAS in Supplementary Table 2). **a**, Distributions of eight summary statistics in the test data and the synthetic data generated by scDesign3 and the four simulators. Each number on top of a violin plot (the distribution of a summary statistic in a synthetic dataset) is the Kolmogorov-Smirnov (KS) distance between the synthetic data distribution (indicated by that violin plot) and the test data distribution. A smaller number indicates better agreement between the synthetic data and the test data in terms of that summary statistic's distribution. **b**, Heatmaps of the gene-gene correlation matrices (showing top 100 highly expressed genes) in the test data and the synthetic data generated by scDesign3 and the four simulators. Pearson's correlation coefficient *r* measures the similarity between two correlation matrices, one from the test data and the other from the synthetic data. **c**, PCA visualization (top two PCs) of the test data and the synthetic data generated by scDesign3 and the four simulators. Colors label cells' pseudotime values; note that only the synthetic data generated by scDesign3 contain the pseudotime truths. An mLISI value close to 2 means that the synthetic data resemble the real data well in the low-dimensional space. **d**, UMAP visualization of the real data and the synthetic data generated by scDesign3 and the four simulators.

https://doi.org/10.1038/s41587-023-01772-1



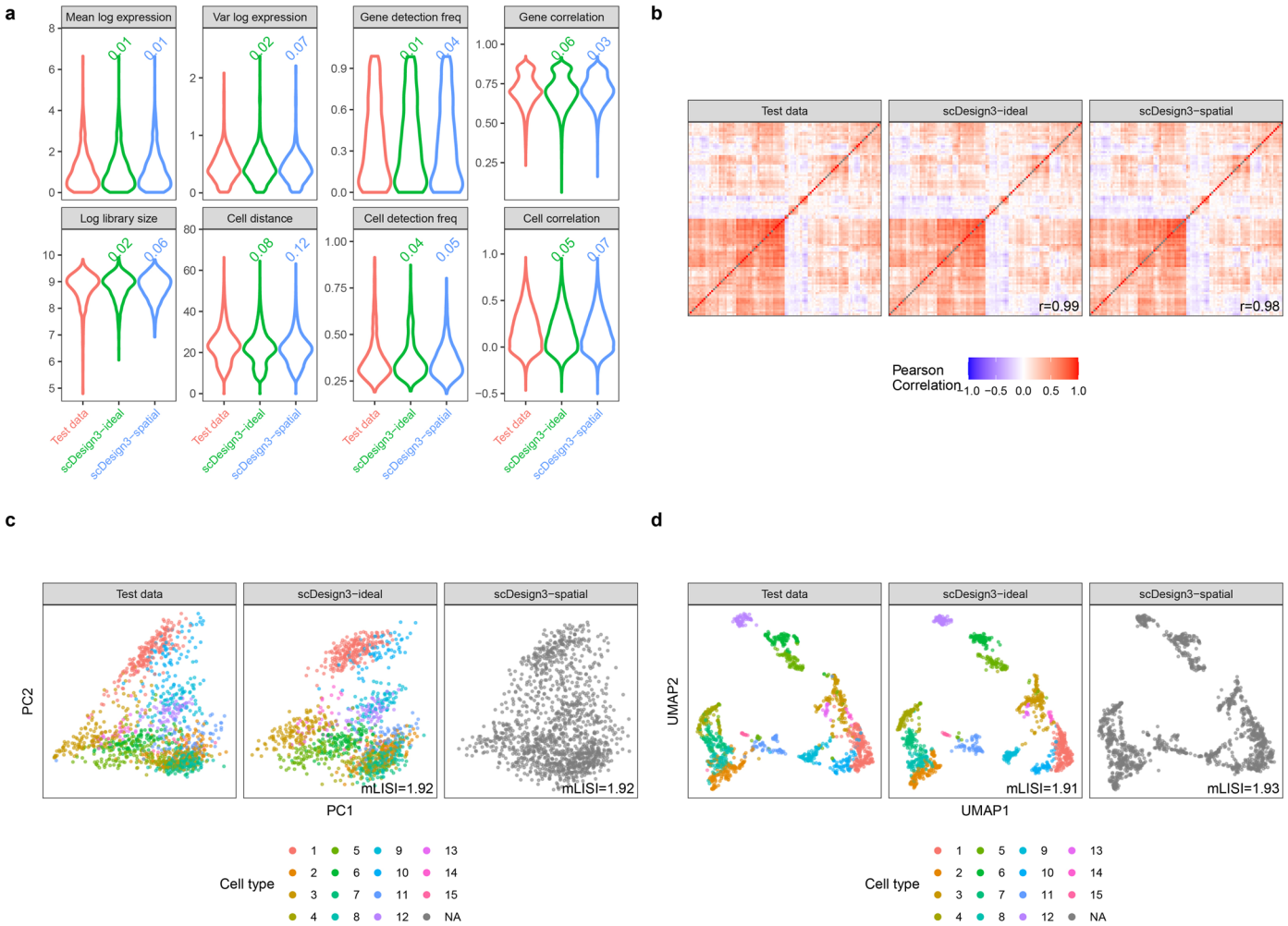
Extended Data Fig. 2 | Benchmarking scDesign3 against four existing scRNA-seq simulators (scGAN, muscat, SPARSim, and ZINB-WaVE) for generating scRNA-seq data from bifurcating trajectories (myeloid progenitors in mouse bone marrow; dataset MARROW in Supplementary Table 2). a, Distributions of eight summary statistics in the test data and the synthetic data generated by scDesign3 and the four simulators. Each number on top of a violin plot (the distribution of a summary statistic in a synthetic dataset) is the Kolmogorov-Smirnov (KS) distance between the synthetic data distribution (indicated by that violin plot) and the test data distribution. A smaller number indicates better agreement between the synthetic data and the test data in terms of that summary statistic's distribution. **b**, Heatmaps of the gene-gene correlation matrices (showing top 100 highly expressed genes) in the test data and the synthetic data generated by scDesign3 and the four simulators. Pearson's correlation coefficient *r* measures the similarity between two correlation matrices, one from the test data and the other from the synthetic data. **c**, PCA visualization (top two PCs) of the test data and the synthetic data generated by scDesign3 and the four simulators. Colors label cells' pseudotime values in two trajectories; note that only the synthetic data generated by scDesign3 contain the pseudotime truths. An mLISI value close to 2 means that the synthetic data resemble the real data well in the low-dimensional space. **d**, UMAP visualization of the real data and the synthetic data generated by scDesign3 and the four simulators.

https://doi.org/10.1038/s41587-023-01772-1



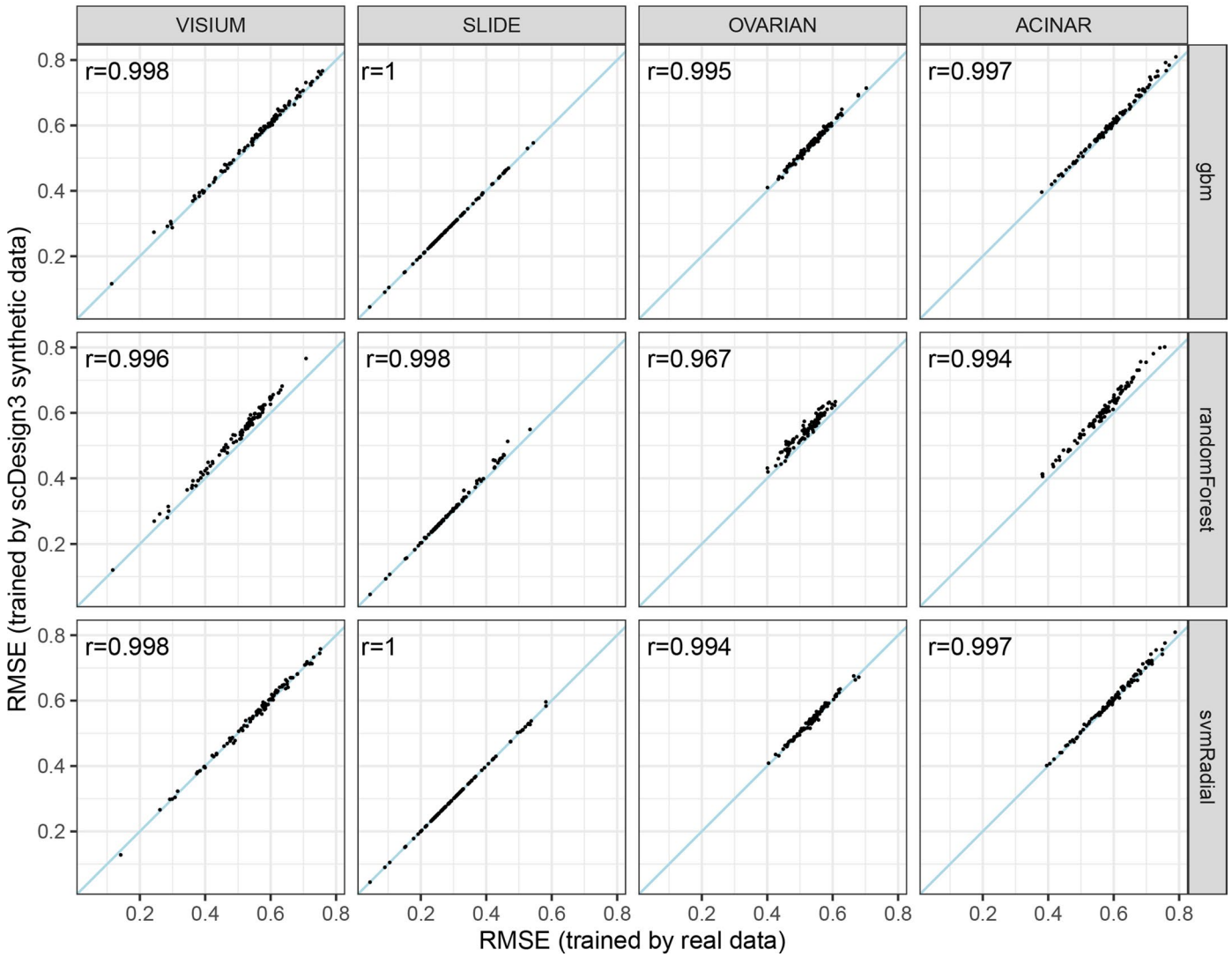
Extended Data Fig. 3 | scDesign3 simulated realistic gene expression patterns in cancer spatial transcriptomics data (datasets OVARIAN and ACINAR in Supplementary Table 2. Human ovarian cancer (**a**) and human prostate cancer, acinar cell carcinoma (**b**). The tissue samples were measured with both H&E (hematoxylin and eosin stain, left) and spatial transcriptomics (right, three cancer-related genes). Large Pearson correlation coefficients (*r*) represent similar spatial patterns in synthetic data and real (test) data.

https://doi.org/10.1038/s41587-023-01772-1



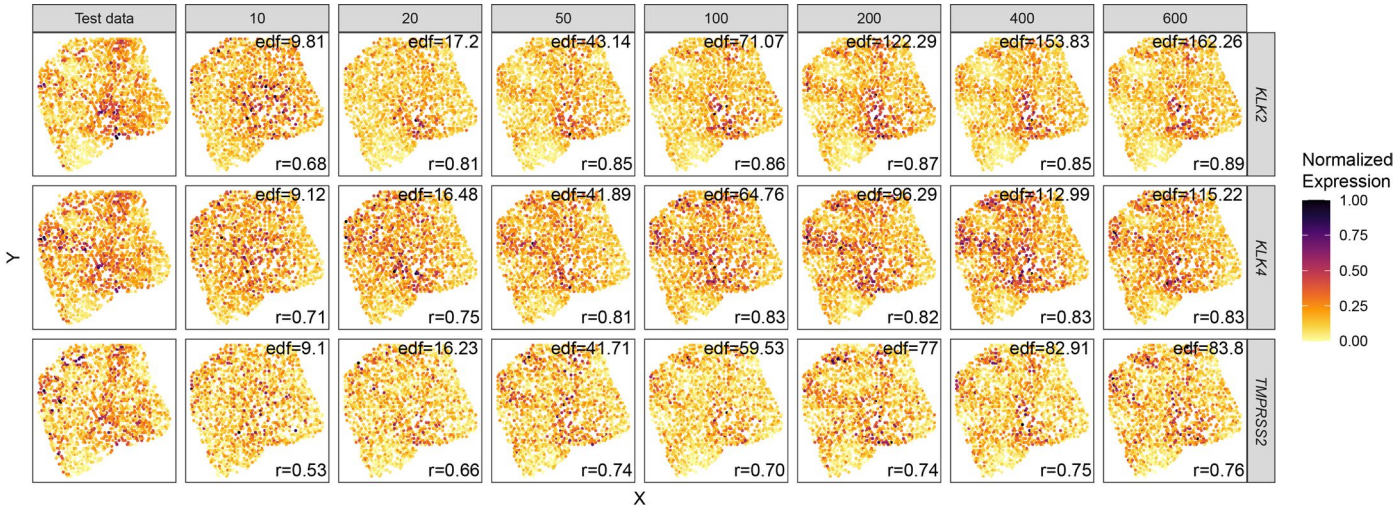
Extended Data Fig. 4 | scDesign3 simulated 10x Visium spatial transcriptomics data (sagital mouse brain slices; dataset VISIUM in Supplementary Table 2). **a**, Distributions of eight summary statistics in the test data and the synthetic data generated by scDesign3 using cell type labels (scDesign3-ideal) and spatial locations (scDesign3-spatial), respectively. Each number on top of a violin plot (the distribution of a summary statistic in a synthetic dataset) is the Kolmogorov-Smirnov (KS) distance between the synthetic data distribution (indicated by that violin plot) and the test data distribution. A smaller number indicates better agreement between the synthetic data and the test data in terms of that summary statistic's distribution. **b**, Heatmaps of the gene-gene correlation matrices (showing top 100 highly expressed genes) in the test data and the synthetic data generated by scDesign3-ideal and scDesign3-spatial. Pearson's correlation coefficient *r* measures the similarity between two correlation matrices, one from the test data and the other from the synthetic data. **c**, PCA visualization (top two PCs) of the real data and the synthetic data generated by scDesign3-ideal and scDesign3-spatial. Cell types (clusters) are labeled by colors. Since the scDesgin3-spatial dataset was based on spatial locations only, it did not contain cell types. An mLISI value close to 2 means that the synthetic data resemble the real data well in the low-dimensional space. **d**, UMAP visualization of the real data and the synthetic data generated by scDesign3-ideal and scDesign3-spatial. In summary, scDesign3 realistically simulated 10x Visium data based on spatial locations without needing cell type annotations.

https://doi.org/10.1038/s41587-023-01772-1



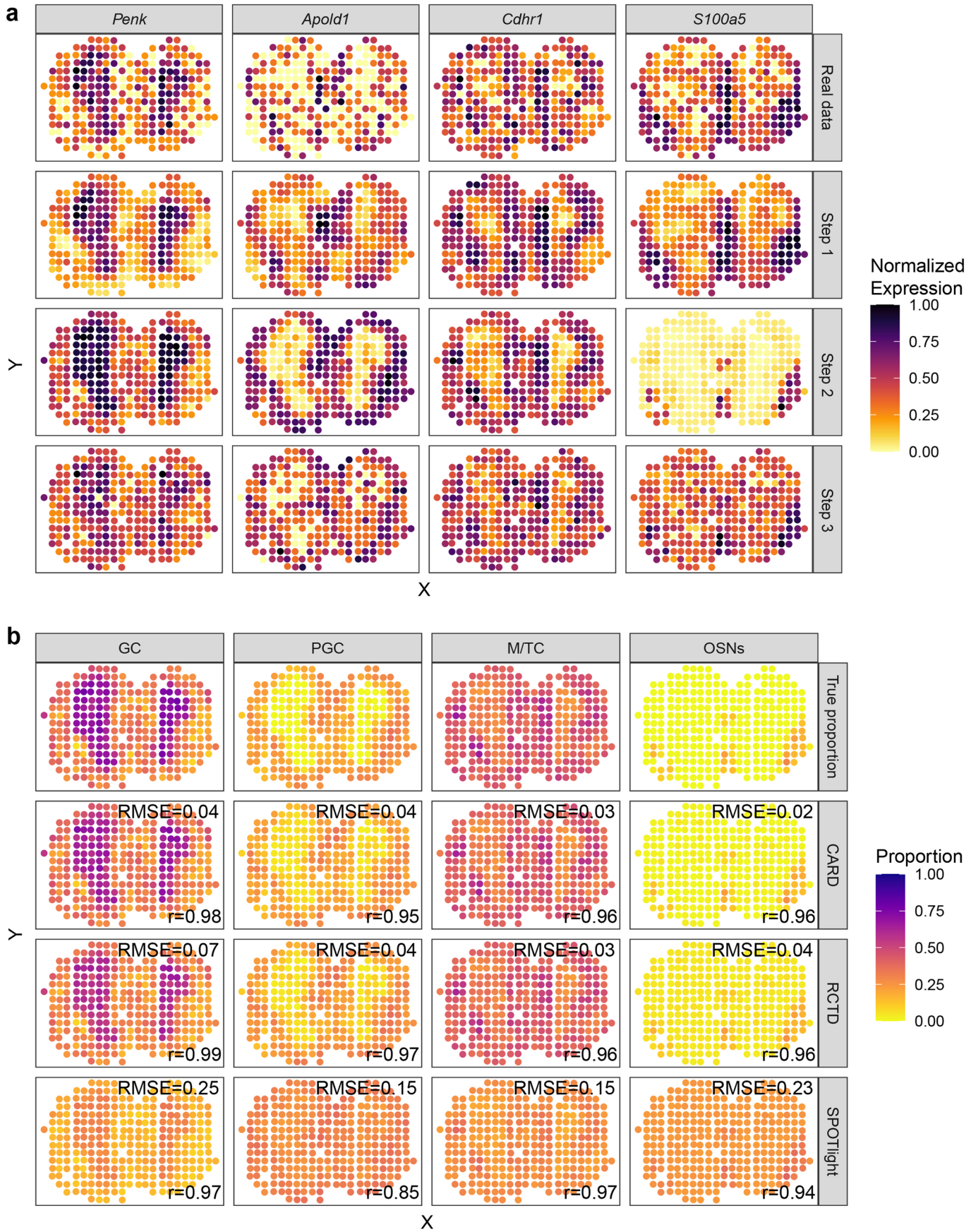
Extended Data Fig. 5 | scDesign3 mimicked spatial transcriptomics data so that prediction algorithms had similar prediction performance when trained on real data or scDesign3 synthetic data. In detail, we first split each of four spatial transcriptomics datasets (VISIUM, SLIDE, OVARIAN, and ACINAR in Supplementary Table 2) into two datasets (training and testing) by randomly splitting the spatial locations into two halves. Second, we used each of the four training datasets to fit scDesign3 and generate the corresponding synthetic dataset. Third, on each pair of training dataset and synthetic dataset (among a total of four pairs), we trained each of three prediction algorithms (gbm: gradient boosting machine; randomForest: random forest; svmRadial: support vector machine with the radial kernel) to predict each gene's expression at a spatial location (input: spatial location; output: the gene's log(count+1) expression level at the location), obtaining a pair of prediction models for each gene. Fourth, we applied each pair of prediction models to the corresponding testing dataset and calculated each model's root-mean-squared error (RMSE) for predicting the corresponding gene, obtaining a pair of RMSEs. As a result, in each panel, we plotted the RMSEs for each prediction algorithm (row) and dataset (column), with each dot in the panel representing a gene. We found all genes' RMSEs highly similar, indicating that scDesign3's synthetic data well mimicked real data.

https://doi.org/10.1038/s41587-023-01772-1



Extended Data Fig. 6 | The effect of *K* on scDesign3's simulation of spatial transcriptomics data (dataset ACINAR in Supplementary Table 2). The rows represent three cancer-related genes; column 1 represents real test data; columns 2–8 represent scDesign3's synthetic data generated using varying *K*, the input basis number. A large Pearson correlation coefficient (*r*) represents similar spatial patterns in synthetic and test data. The effective degrees of freedom (edf) represents the wiggliness of the fitted surface. With a larger *K*, scDesign3 can fit more complex patterns. The overfitting issue is accounted for by the automatic smoothness estimation[39]: when *K* is sufficiently large, edf (model complexity) and *r* (model goodness-of-fit) both become stable.

https://doi.org/10.1038/s41587-023-01772-1

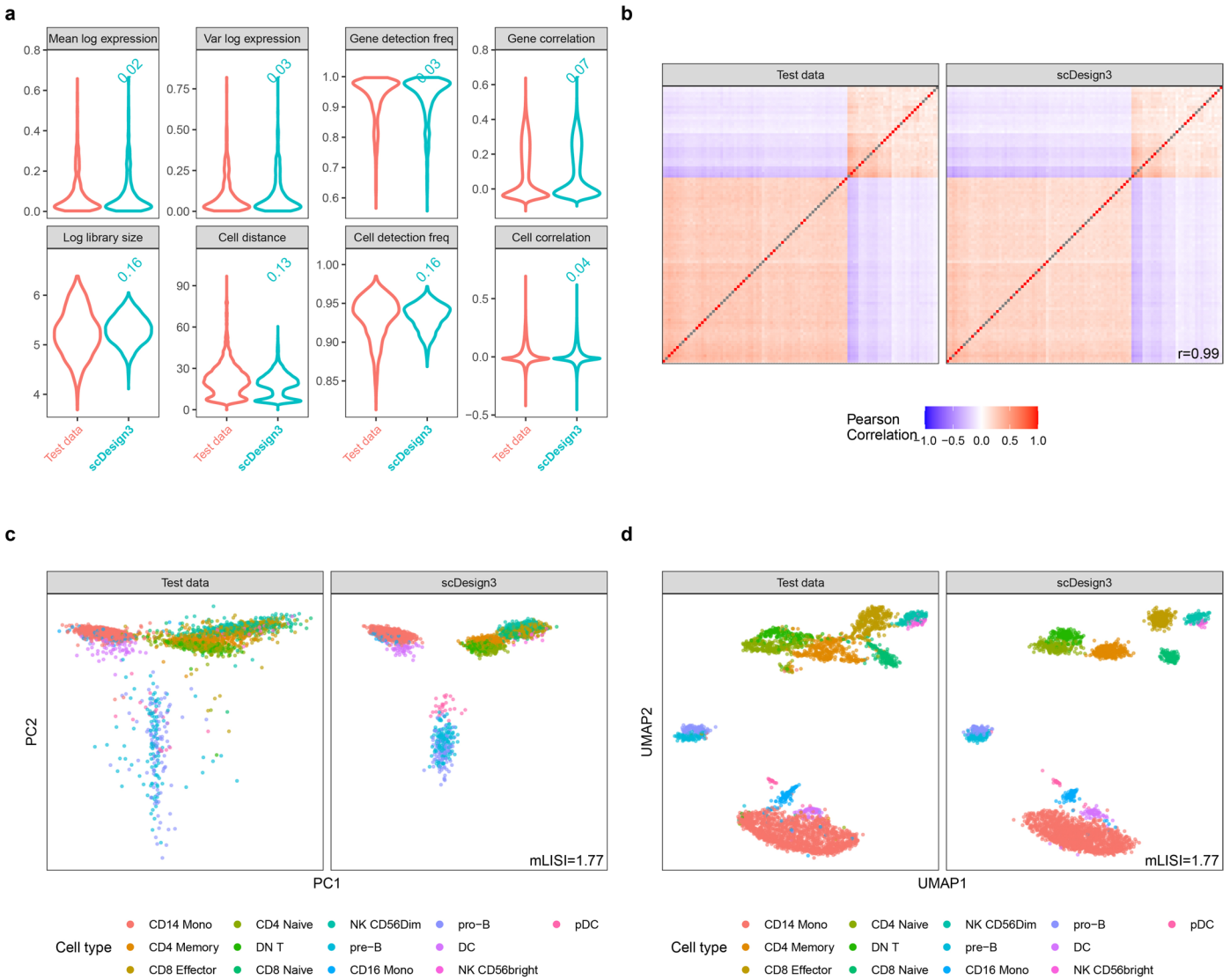


Extended Data Fig. 7 | See next page for caption.

https://doi.org/10.1038/s41587-023-01772-1

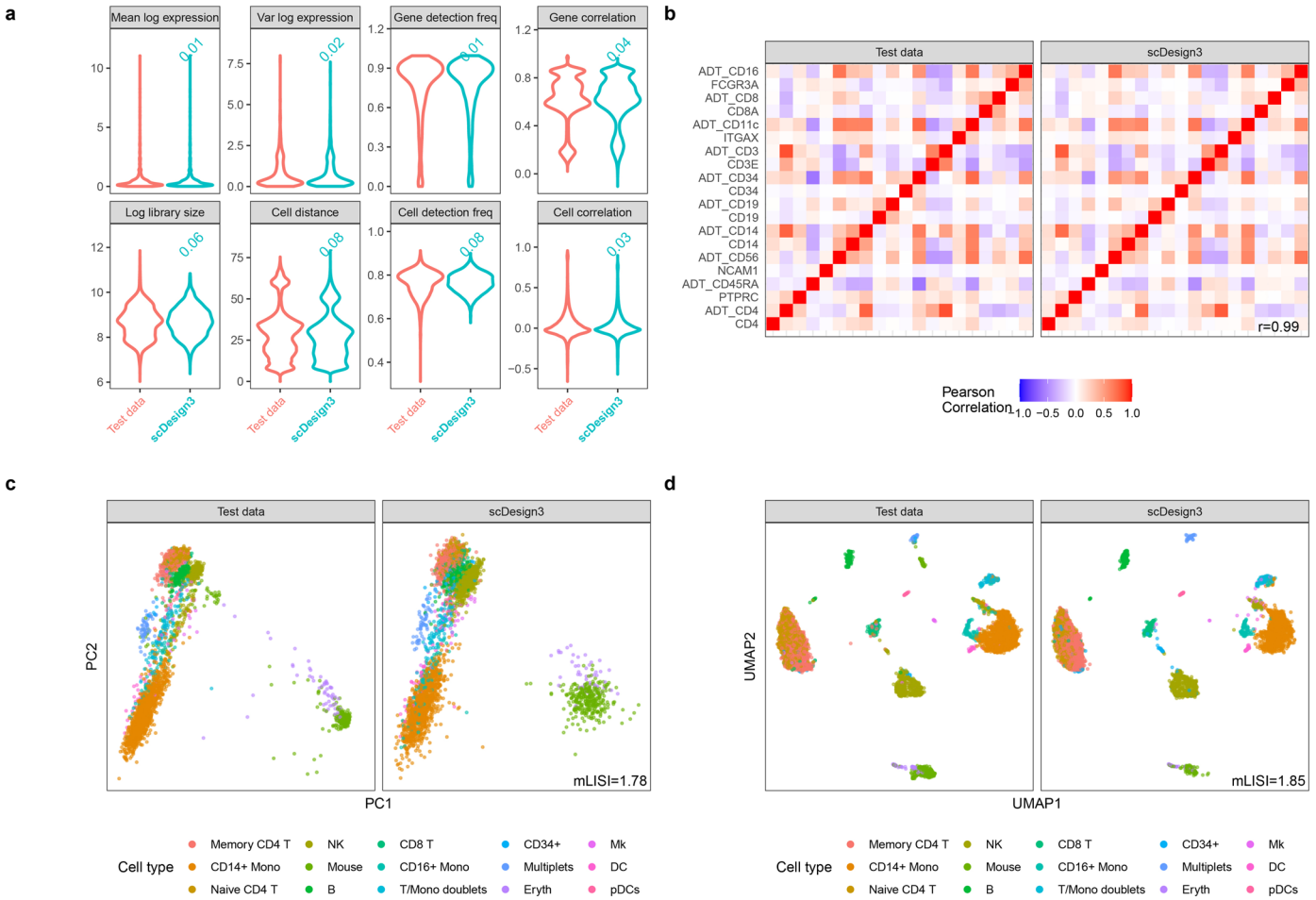
Extended Data Fig. 7 | scDesign3 simulated spot-resolution spatial transcriptomics data for benchmarking cell-type deconvolution algorithms (datasets MOB-SP and MOB-SC in Supplementary Table 2). a, scDesign3's synthetic spot-resolution data well mimicked real data (top row), showing similar expression patterns for four cell-type marker genes (columns). scDesign3 used three steps to generate the spot-resolution data. Step 1: every gene's estimated mean expression level at each spot (as a smooth function of spot location) by scDesign3. Step 2: every gene's predicted expression level at each spot from CIBERSORT's estimated cell-type proportions at the spot (considered as the 'true proportions') and the gene's cell-type-specific expression levels (from the reference scRNA-seq data). Step 3: every gene's simulated expression level at each spot by scDesign3 (from the true cell-type proportions at the spot and scDesign3's synthetic scRNA-seq data). **b**, Using scDesign3 synthetic data, we benchmarked three spatial cell-type deconvolution algorithms (CARD[6], RCTD[7], and SPOTlight[8]). For each of the four cell types (columns), we used two metrics-Pearson correlation (*r*) and root-mean-square error (RMSE)-to compare the proportions estimated by each deconvolution algorithm (rows 2-4) to the true proportions (top row). Large *r* values represent similar spatial patterns of proportions, while small RMSE values represent similar values of proportions. Although all three algorithms well captured the spatial patterns of each cell type's proportions (evidenced by large *r* values), CARD and RCTD outperformed SPOTlight by estimating cell-type proportions more accurately (evidenced by smaller RMSE values).

https://doi.org/10.1038/s41587-023-01772-1



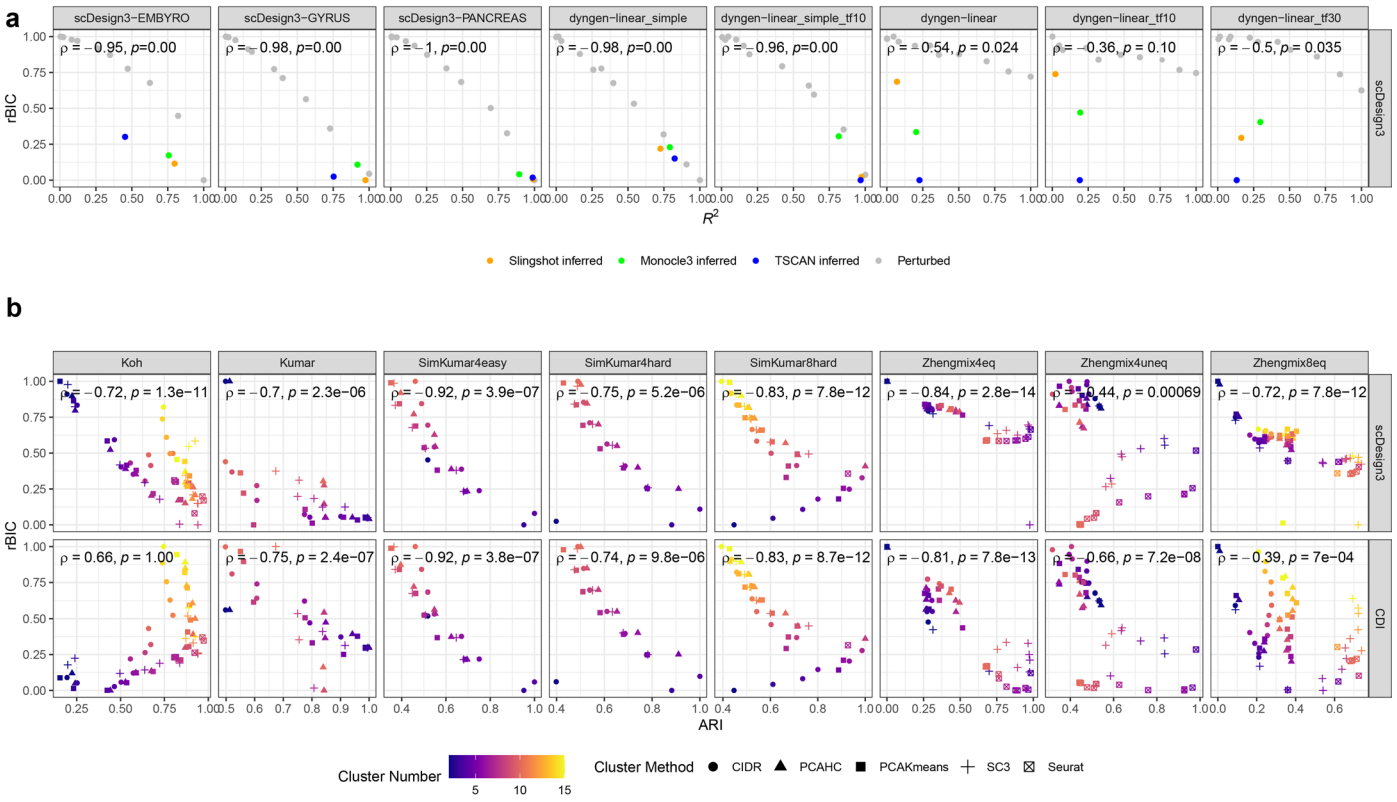
Extended Data Fig. 8 | scDesign3 simulated scATAC-seq data (human PBMCs; dataset ATAC in Supplementary Table 2). a, Distributions of eight summary statistics in the test data and the synthetic data generated by scDesign3 using cell type labels. Each number on top of a violin plot (the distribution of a summary statistic in a synthetic dataset) is the Kolmogorov-Smirnov (KS) distance between the synthetic data distribution (indicated by that violin plot) and the test data distribution. A smaller number indicates better agreement between the synthetic data and the test data in terms of that summary statistic's distribution. **b**, Heatmaps of the peak-peak correlation matrices in the test data and the synthetic data generated by scDesign3. Pearson's correlation coefficient *r* measures the similarity between two correlation matrices, one from the test data and the other from the synthetic data. **c**, PCA visualization (top two PCs) of the test data and the synthetic data generated by scDesign3. Cell types are labeled by colors. An mLISI value close to 2 means that the synthetic data resemble the test data well in the low-dimensional space. **d**, UMAP visualization of the test data and the synthetic data generated by scDesign3.

https://doi.org/10.1038/s41587-023-01772-1



Extended Data Fig. 9 | scDesign3 simulated CITE-seq data (human PBMCs; dataset CITE in Supplementary Table 2). a, Distributions of eight summary statistics in the test data and the synthetic data generated by scDesign3. The CITE-seq dataset contains simultaneous measurements of each cell's gene expression and surface protein abundance captured by Antibody-Derived Tags (ADTs). Each number on top of a violin plot (the distribution of a summary statistic in a synthetic dataset) is the Kolmogorov-Smirnov (KS) distance between the synthetic data distribution (indicated by that violin plot) and the test data distribution. A smaller number indicates better agreement between the synthetic data and the test data in terms of that summary statistic's distribution. **b**, Heatmaps of the gene and protein correlation matrices (10 proteins with names starting with 'ADT' and their corresponding genes) in the test data and the synthetic data generated by scDesign3. Pearson's correlation coefficient *r* measures the similarity between two correlation matrices, one from the test data and the other from the synthetic data. scDesign3 preserved the correlations between the RNA and protein expression levels of the 10 surface proteins. **c**, PCA visualization (top two PCs) of the test data and the synthetic data generated by scDesign3. Cell types are labeled by colors. An mLISI value close to 2 means that the synthetic data resemble the real data well in the low-dimensional space. **d**, UMAP visualization of the test data and the synthetic data generated by scDesign3.

https://doi.org/10.1038/s41587-023-01772-1



Extended Data Fig. 10 | scDesign3 provides unsupervised measures of the goodness-of-fit of pseudotime, clusters, and inferred spatial locations. For visual clarity, we plot the relative BIC or AIC (rBIC or rAIC) by re-scaling scDesign3's marginal BIC or AIC to [0, 1]. **a**, The scDesign3 rBIC (unsupervised) is negatively correlated with the R^2 (supervised). Each R^2 was calculated between the set of perturbed or inferred pseudotimes and the set of true pseudotimes in each of the eight datasets (the column names). The P value is from the one-sided test of Spearman's rank correlation ρ. The true pseudotime is the ground truth used for generating the synthetic data. **b**, Comparison of the scDesign3 rBIC and the Clustering Deviation Index (CDI) rBIC (rescaled to [0, 1])[33]. The color scale shows the number of clusters, and the shapes represent clustering algorithms. We found the scDesign3 rBIC (unsupervised) negatively correlated with the ARI (supervised). The P value is from the one-sided test of Spearman's rank correlation ρ. We also found the scDesign3 rBIC to perform better or similarly to the CDI on six out of the eight datasets (the column names). **c**, The scDesign3 rAIC (unsupervised) is negatively correlated with the mean cosine similarity (supervised). The mean cosine similarity was calculated between the set of perturbed or inferred locations and the set of true locations in each of the two spatial datasets (the column names). The P value is from the one-sided test of Spearman's rank correlation ρ. The true locations are the ground truth used for generating the semi-synthetic data. Due to the high complexity of spatial patterns, the scDesign3 rAIC (left) outperformed the scDesign3 rBIC (right) for penalizing the model complexity less.

nature portfolio

Corresponding author(s): Jingyi Jessica Li

Last updated by author(s): Mar 18, 2023

Reporting Summary

Nature Portfolio 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 Portfolio 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	All datasets are downloaded within by using R (version 4.1.0 - 4.2.1) scripts. The codes for downloading data can be found in Zenodo: https://doi.org/10.5281/zenodo.7110762 with files named with "_exploration". The cell-type data for clustering goodness-of-fit is from R package DuoClustering2018 (version 1.10.0).
Data analysis	Analysis is performed in R (version 4.1.0 - 4.2.1). The R packages for bioinformatics analysis include: scran (version 1.20.1), Signac (version 1.7.0), Seurat (version 4.1.1), Slingshot (version 2.2.1), CellMixS (version 1.8.0), dyngen (version 1.0.3), muscat (version 1.6.0), SPARSim (version 0.9.5), zinbwave (version 1.15.3). The R packages for pseudotime inference are slingshot (version 2.4.0), monocle3 (version 1.0.0) and TSCAN (version 1.34.0). The softwares for spatial location inference are Seurat (R version 4.1.1), Tangram (Python version 1.0.0), and novoSpaRc (Python version 0.4.3). The R packages for spatial data deconvolution include CARD (version 1.0), spacexr (version 2.1.6), and SPOTlight (version 1.0.1). The R package for prediction is caret (version 6.0-93). The Python module for multi-omics integration is Pamona (version 0.1.0). The R packages for general analysis and visualization include: irlba (version 2.3.5), umap (version 0.2.8.0), ggplot2 (version 3.3.6), ggpubr (version 0.4.0). The scDesign3 is available at: https://github.com/SONGDONGYUAN1994/scDesign3. The scDesign3 version in the study is 0.99.0. The R packages for statistical modeling in scDesign3 are mgcv (version 1.8-40), gamlss (version 5.4-3), rvinecoplib (version 0.6.2.1.1), Rfast (version 2.0.6), and stats (version 4.4.2). The read coverage plot is generated by IGV (version 2.12.3). The scGAN is downloaded from https://github.com/imsb-uke/scGAN. The code for analysis can be found in: https://doi.org/10.5281/zenodo.7110761 .

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 Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All datasets used in the study are publicly available. Supplementary Table S2 lists the datasets from 19 published studies and their original sources. The preprocessed datasets are available at: https://doi.org/10.5281/zenodo.7110761 .

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	Not applicable.
Population characteristics	Not applicable.
Recruitment	Not applicable.
Ethics oversight	Not applicable. All datasets are from published studies.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

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	We apply scDesign3 on real datasets from 19 published studies. The cell numbers of each dataset are reported in Supplementary Table S2. We decide the number of datasets used based on the following rationale: for data simulation part, we use at least two datasets for one major simulation direction (e.g., trajectory, spatial, chromatin accessibility) to make sure that our simulator works for different experimental protocols and biological cases. For interpretation of parameters and alteration of parameters, we use one dataset for one application since the main purpose is to illustrate the usage of scDesign3. For assessing the goodness-of-fit, we use >= 2 datasets for each latent variable type (e.g., cell type, pseudotime, spatial) to make sure this metric works for diverse datasets.
Data exclusions	Some cell-level filtering are performed by pipelines as their default settings (Seurat, Signac, scran).
Replication	Replication is not relevant to this study because we do not design/perform repeat experiments on technical/biological replicates by ourselves; all datasets are public datasets with the labels (e.g., replicates if any) from the original studies.
Randomization	For each real dataset used as reference, we randomly split it into half training data and half test data to avoid overfitting. The code and reproducible seed can be found in Zenodo: https://doi.org/10.5281/zenodo.7110761.
Blinding	Blinding is not relevant to this study because we do not design/perform experiments by ourselves; all datasets are public datasets with the labels (e.g., conditions) from their original studies.

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
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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