

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
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| ☒ | ☐ | 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
<i>Give P values as exact values whenever suitable.</i> |
| ☒ | ☐ | 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 scanpy==1.7.2 seaborn==4.1.0

Data analysis Custom software: <https://github.com/TencentAILabHealthcare/scBERT>
Public softwares: scipy==1.5.4 torch==1.8.1 numpy==1.19.2 pandas==1.1.5 scanpy==1.7.2 scikit-learn==0.24.2 transformers==4.6.1.

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 data used in this study are publicly available and the usages are fully illustrated in the Method section. The published Panglao dataset was downloaded from the PanglaoDB website (<https://panglao.db.se/>). The published Zheng68k dataset was downloaded from the "Fresh 68K PBMCs" on the website (<https://support.10xgenomics.com/single-cell-gene-expression/datasets> (SRP073767)). The published pancreatic datasets were downloaded from the github (<https://>

hemberg-lab.github.io/scRNA.seq.datasets/ (Baron: GSE84133, Muraro: GSE85241, Segerstolpe: E-MTAB-5061, Xin: GSE81608)). The MacParland dataset was downloaded from the repository (<https://www.ncbi.nlm.nih.gov/geo/> (GSE115469)). The heart datasets were downloaded from the websites (<https://data.humancellatlas.org/explore/projects/ad98d3cd-26fb-4ee3-99c9-8a2ab085e737> and https://singlecell.broadinstitute.org/single_cell/study/SCP498/transcriptional-and-cellular-diversity-of-the-human-heart). The lung dataset for COVID-19 study was downloaded from the website (<https://doi.org/10.6084/m9.figshare.11981034.v1>). The adult human cell atlas of 15 major organs dataset was downloaded from the website (<https://www.ncbi.nlm.nih.gov/geo/> (GSE159929)).

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender	N/A. Didn't perform any sex- or gender- based analyses.
Population characteristics	N/A. Didn't conduct any population based analyses.
Recruitment	Used publically available data. Didn't recruit any participants.
Ethics oversight	Used publically available data. No study protocol approval is needed.

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](https://www.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	In this work, we investigated the model performance on nine datasets: (1) Zheng68k dataset contains 68450 cells; (2) Baron dataset contains 8569 cells; (3) Xin dataset contains 1449 cells; (4) Segerstolpe dataset contains 2133 cells; (5) Muraro dataset contains 2122 cells; (6) MacParland dataset contains 8444 cells; (7) Tucker dataset contains 287269 cells; (8) Lung dataset contains 39778 cells; (9) Human Cell Atlas dataset contains 84363 cells. All the nine datasets are publicly available and illustrated in the data availability statement. The datasets are used in different experiments with elaborated experiment design in the method section.
Data exclusions	No data was excluded from the analysis.
Replication	In this work, we validated the model performance using five-fold cross-validation. We therefore use the average values from different validation sets to draw the benchmarking boxplots. The cross-organ and cross-cohort experiments are performed using different combinations of training sets and test sets.
Randomization	For pre-training on reconstruction task, we randomly mask the gene expression of non-zero values with 15% masking ratio.
Blinding	During the model pre-training on unlabeled PanlaoDB dataset, all the cell type labels are unseen for the model. All the nine datasets for benchmarking are not included in the dataset for pre-training. As for the novel cell type discovery experiments, the novel cell types are unseen during fine-tuning stage. Therefore, potential information leakage or overfitting was avoided during the pre-training and fine-tuning.

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	Involvement in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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