nature portfolio

Corresponding author(s): Yuedong Yang

Last updated by author(s): Apr 3, 2025

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 data analyzed within this manuscript are publicly available. No additional software was used for the data collection process.
Data analysis	We performed the data analysis using our developed CellFM software (Version: 1.0.0) in this paper, which is available at the GitHub repository: [https://github.com/biomed-AI/CellFM]. The following software packages were used in our code: python v3.9.0, numpy v1.24.4, scikit-learn v1.3.0, scipy v1.9.3, mindspore v2.2.10, tqdm v4.65.0, h5py v3.10.0, scanpy v1.10.1, anndata v0.10.5, pandas v1.5.3, matplotlib v3.6.3, louvain v0.8.1, leidenalg v0.10.2, Genecompass v1.0.0, UCE v1.0.0, scBert v1.0.0, scFoundation v1.0.0, scmap v1.0.0, CellOTv1.0.0, and GEARS v1.0.0, and SVM v1.0.0.

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

The training datasets utilized for CellFM are comprehensively documented in Supplementary_Metadata_information.xlsx.

The evaluation datasets utilized in this study are publicly available through the original publications cited in our manuscript. Detailed access links and dataset descriptions are provided in Supplementary Table S4 for reference. Additionally, we have summarized all datasets as follows:

The pre-training datasets can be retrieved from the NCBI GEO, ENA, GSA, and ImmPort.

For the gene function prediction task, The datasets used in T1, T2, and T3 tasks were downloaded from the website: https://panglaodb.se/view_data.php?sra=SRA553822&srs=SRS2119548; The gene labels of these three tasks were downloaded from the website: https://huggingface.co/datasets/ctheodoris/Genecorpus-30M/tree/main/example_input_files/gene_classification/bivalent_promoters.

For the perturbation prediction task, the Norman and Adamson datasets were retrieved from the GEARs API via the folowing links (https://dataverse.harvard.edu/api/access/datafile/6154020; https://dataverse.harvard.edu/api/access/datafile/6154417).

For the cell annotation task, the Myeloid dataset is publicly accessible from the Gene Expression Omnibus (GE0) database using the accession number GSE154763. The datasets including Pancrm, HumanPBMC, Cell Lines, MCA, and DC were downloaded from the link: https://github.com/JinmiaoChenLab/Batch-effect-removal-benchmarking/tree/master. The PBMC 368K dataset was downloaded from the link : https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc3k;https://www.10xgenomics.com/resources/datasets/fresh-68-k-pbm-cs-donor-a-1-standard-1-1-0. The Immune dataset was downloaded from the link: https://figshare.com/articles/dataset/Benchmarking_atlas-level_data_integration_in_single-cell_genomics_integration_task_datasets_Immune_and_pancreas_/12420968. The processed hPancreas datasets retrieved from https:/github.com/JackieHanlab/TOSICA. The Liver dataset was downloaded from the link: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE115469;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE124395. The Lung, PBMC, Skin, and Heart datasets were derived from studies corresponding to the following DOIs: https://doi.org/10.1038/s41592-021-01336-8, https://doi.org/10.1101/2024.04.02.58782, and https://doi.org/10.48550/arXiv.2408.12373, respectively. The human brain cell atlas was from DOI: 10.1126/science.add7046 and the datasets Tabula Sapiens including two versions were from DOI: 10.1126/science.abl4896 and doi.org/10.1101/2024.12.03.626516, respectively. The datasets BCC_GSE123813 and LIHC_GSE140228 were downloaded from http://tisch.comp-genomics.org/gallery/.

For Gene Network Analysis, the processed Immune Human dataset was accessed from https:/doi.org/10.6084/m9.figshare.1242096852. the Adamson datasets was retrieved from the GEARs API via the links https://dataverse.harvard.edu/api/access/datafile/6154417.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

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	To ensure comprehensive pre-training, we compiled a large-scale dataset from multiple web-based resources. Our methodological approach and sample selection followed well-established protocols from previous studies, thereby eliminating the need for additional sample size calculations. To comprehensively evaluate CellFM's capabilities, we utilized exclusively publicly available datasets, including: Pancrm, HumanPBMC, Immune, MCA, Cell Lines, PBMC 368K, DC, Myeloid, hPancreas (train), hPancreas (test), Liver, Lung, PBMC, Skin, Heart, LIHC_GSE140228, and BCC_GSE123813. These datasets were carefully selected to enable rigorous evaluation of classification performance across diverse conditions, including batch effects and cell subtype discrimination. For assessing CellFM's data integration capabilities, we employed PBMC 10K, The Human Brain Cell Atlas, Tabula Sapiens V1, and Tabula Sapiens V2 datasets. The perturbation prediction performance was evaluated using the Norman and Adamson datasets. All datasets have been previously validated in benchmark studies such as scGPT [1] and scFoundation [2], ensuring their statistical robustness for methodological benchmarking. [1] Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., & Wang, B. (2024). scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nature Methods, 21(8), 1470-1480. [2] Hao, M., Gong, J., Zeng, X., Liu, C., Guo, Y., Cheng, X., ... & Song, L. (2024). Large-scale foundation model on single-cell transcriptomics. Nature methods, 21(8), 1481-1491.
Data exclusions	We performed quality control of scRNA-seq data based on the common used and pre-established criteria in this field
Replication	Due to potential slight variations in outcomes when the random seed is not fixed during the execution of the deep learning model, we conducted ten runs to assess its robustness.
Randomization	Randomization is not relevant to this study since we do not collect any new experimental data
Blinding	Since all the data used in this study have been previously published, it is not feasible to conduct blind investigations during the reanalysis of the data. However, the allocation information is blinded to the computational algorithms during the performance evaluation.

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
☒	☐ Plants

Methods

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

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A