nature portfolio

Corresponding author(s): DAPR NBT-BC69723

Last updated by author(s): 07/22/2026

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	No software was used for data collection
Data analysis	The source code is available on GitHub at https://github.com/shiftbioscience/Perturbation-Models-Outperform-Baselines We used the following Python packages: fastcluster>=1.2.6 igraph>=0.11.8 ipdb>=0.13.13 ipython>=5.2.0 leidenalg>=0.10.2 nbconvert>=7.16.6 scanpy>=1.11.1 scienceplots>=2.1.1 scipy (==1.14.1) seaborn>=0.13.2 ipykernel (>=6.29.5,<7.0.0) numpy (==2.2.0) pandas (==2.2.2) tqdm==4.66.5 jupytext>=1.17.1

hydra-core>=1.3.0
docker>=7.0.0
omegaconf>=2.3.0
matplotlib>=3.5.0
pytest>=7.0.0
upsetplot>=0.9.0
openai>=1.99.9
pybiomart>=0.2.0
requests>=2.32.3
mkdocs>=1.6.1
pymdown-extensions>=10.16.1
mkdocs-material>=9.6.17
mkdocstrings-python>=1.17.0
nvitop>=1.5.3
scperturb>=0.1.0

Using the UV python environment manager

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

Dataset Link

Adamson16	https://zenodo.org/record/10044268/files/AdamsonWeissman2016_GSM2406681_10X010.h5ad
Frangieh21	https://zenodo.org/record/7041849/files/FrangiehIzar2021_RNA.h5ad
Kaden25fibroblast	https://zenodo.org/records/15200179/files/fibroblast_CRISPRa_final_pop.h5ad?download=1
Kaden25rpe1	https://zenodo.org/records/15213619/files/RPE1_CRISPRa_final_population.h5ad?download=1
Nadig25hepg2	https://ftp.ncbi.nlm.nih.gov/geo/series/GSE264nnn/GSE264667/suppl/GSE264667%5Fhepg2%5Fraw%5Fsinglecell%5F01.h5ad
Nadig25jurkat	https://ftp.ncbi.nlm.nih.gov/geo/series/GSE264nnn/GSE264667/suppl/GSE264667%5Fjurkat%5Fraw%5Fsinglecell%5F01.h5ad
Norman19	https://zenodo.org/records/7041849/files/NormanWeissman2019_filtered.h5ad?download=1
Replogle22k562	https://zenodo.org/record/10044268/files/ReplogleWeissman2022_K562_essential.h5ad
Replogle22k562gwps	https://zenodo.org/record/10044268/files/ReplogleWeissman2022_K562_gwps.h5ad
Replogle22rpe1	https://zenodo.org/record/10044268/files/ReplogleWeissman2022_rpe1.h5ad
Sunshine23	https://zenodo.org/record/10044268/files/SunshineHein2023.h5ad
Tian21crispra	https://zenodo.org/record/10044268/files/TianKampmann2021_CRISPRa.h5ad
Tian21crispri	https://zenodo.org/record/7041849/files/TianKampmann2021_CRISPRi.h5ad
Wessels23	https://zenodo.org/records/13350497/files/WesselsSatija2023.h5ad?download=1

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	We did not perform any sex-based analyses
Reporting on race, ethnicity, or other socially relevant groupings	We did not perform any race-based analyses
Population characteristics	See above
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

Life sciences study design

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

Sample size	We used published datasets, so sample size was defined by the studies that generated the data and any relevant post-processing. For the unseen single perturbation task, we perform 5-fold cross-validation, using a 70/10/20% train/validation/test split in each fold. For the unseen combo perturbation task, we perform 2-fold cross-validation over the set of combination perturbations.
Data exclusions	For every dataset, we filter out any perturbation with less than 12 measured cells, any cell with less than 200 expressed genes, and any gene expressed in less than 3 cells. To balance the datasets, we enforce a maximum number of cells per perturbation calculated as the mean of cells per perturbation after previous filters were applied. We further select a distinct gene set per dataset composed by the union of the top 8192 highly variable genes plus all perturbed genes. The only exception to the above protocol is Replogle22k562wps in which we randomly subsampled perturbations to 2,500 due to memory constraints.
Replication	We provide the code to reproduce our results at https://github.com/shiftbioscience/Perturbation-Models-Outperform-Baselines. However, we did not attempt to independently replicate our results.
Randomization	Not required as we were not running any new gene perturbation experiments. Controlling covariates is not applicable to the study because we do not generate new data and we benchmark existing computational methods as they were originally designed.
Blinding	Not required as we were not running any new gene perturbation experiments.

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