nature portfolio

Corresponding author(s): Jonas Fuxe

Last updated by author(s): Mar 20, 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	NIS elements software (version AR 5.30.02)
Data analysis	CellProfiler (version 4.2.5); Python (version 3.12.4); R software (4.4.0); GraphPad Prism (version 10.4.0)

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

Image data used to perform analyses in the manuscript are available at https://figshare.scilifelab.se/account/home#/projects/234446;
The codes, pipelines, and datasets used are available in the following GitHub repository: https://github.com/jonfux/Cell-Painting-EMT

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	NA
Reporting on race, ethnicity, or other socially relevant groupings	NA
Population characteristics	NA
Recruitment	NA
Ethics oversight	NA

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	The multi-level sampling structure was intentionally designed to capture both biological and technical variability. Biological variability is accounted for by using multiple independent plates, while technical variability is addressed by including multiple wells and images per well. Due to our sample size design the statistical power of our study allowed for detection of subtle phenotypic differences with high confidence.
Data exclusions	Aggregated data from the three plates were preprocessed to remove missing values and normalized per plate using robust scaling based on the median and interquartile range (IQR) per feature across conditions. Feature selection was applied to remove highly correlated (>0.9), low-variance (<0.01), and metadata-rich columns using Pycytominer (version 1.1.0). A blocklist was applied to remove specific features from the normalized and aggregated dataset. Features were excluded based on previous experience and predefined patterns associated with irrelevant or redundant measurements.
Replication	All attempts to replicate the findings were succesful
Randomization	Treatment allocation was done to ensure that replicates were assigned different spots in the 96-well plates. For the in silico analysis, the dataset was randomly divided into training and test sets to enable rigorous training and evaluation of multiple classifiers.
Blinding	The investigator seeded cells in the wells but the machine learning was performed using label-blinding.

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

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	ATCC
Authentication	ATCC
Mycoplasma contamination	Regularly checked at Eurofins
Commonly misidentified lines (See ICLAC register)	None of the used ones

Plants

Seed stocks	NA
Novel plant genotypes	NA
Authentication	NA