nature portfolio

Corresponding author(s): Fatemeh Vafaee, Marcel E. Dinger, Emily Oates

Last updated by author(s): Mar 10, 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	Data were retrieved directly from their respective host servers utilizing the wget command in Bash.
Data analysis	The analysis was conducted using Python 3.11.0, which can be accessed at Python's official website (https://www.python.org/). The following packages, essential for the analysis, were obtained and installed via pip from PyPI (Python Package Index, https://pypi.org/). anndata==0.10.5.post1 axial_positional_embedding==0.2.1 Bio==1.6.2 biopython==1.82 data==0.4 dna==0.0.1 einops==0.7.0 eval==0.0.5 h5py==3.9.0 local_attention==1.8.6 matplotlib==3.7.1 numpy==1.25.0 openpyxl==3.1.2 packaging==21.3 pandas==2.0.2 pySankey==0.0.1

scikit_learn==1.2.2
scipy==1.12.0
seaborn==0.13.2
statsmodels==0.14.0
torch==2.0.1
train==0.0.5
transformers==4.32.1
utils==1.0.2

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 supervised aligned gene embeddings derived from the protein-coding genes-only GPT model, as well as from the model of coding genes, lncRNAs + pseudogenes, are available. These embeddings allow for straightforward calculation of pairwise similarity. The 37,926 by 37,926 pairwise similarity matrix, produced by averaging the similarities from the supervised alignment approach and those from shared embeddings + supervised alignment approach, is also available. Provided files are made available under the Creative Commons Attribution Non Commercial 4.0 International License. All data is free to use for non-commercial purposes.

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	Sex and gender were not variables in the study design, the study focused on gene expression data and did not differentiate or analyze the data based on sex or gender.
Reporting on race, ethnicity, or other socially relevant groupings	The study did not utilize socially relevant categorization variables such as race or ethnicity. The primary focus was on gene expression data and its AI modeling.
Population characteristics	The samples represented a wide spectrum of phenotypes, tissues, cell lines, ages, and genders. The study did not explicitly control for covariates such as age, sex, or ethnicity in its analysis.
Recruitment	The research was a secondary analysis of publicly available data from the ARCHS4 database.
Ethics oversight	All data collection and processing were conducted in compliance with ethical standards and procedures, with oversight and approval by the University of New South Wales (UNSW) Human Research Ethics Committee.

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	No statistical method was used to predetermine sample size.
Data exclusions	Samples were selected based on criteria ensuring high quality: total gene expression read counts exceeding one million, a minimum of 2000 genes with non-zero expression, and a single-cell probability below 0.5. Duplicate gene symbols were handled by retaining those with the highest mean expression. Genes with mean normalized expression above 0.1 were maintained.

Replication	The study did not involve direct replication of experiments but focused on the analysis of existing large-scale transcriptome data.
Randomization	The study did not involve randomization. However, for the BERT model, 15% of genes with non-zero expression had either their identity (gene symbol) or expression values masked randomly for model training.
Blinding	The study did not involve blinding, as it primarily dealt with computational modeling and analysis of pre-existing transcriptomic data.

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	No seed stock was used.
Novel plant genotypes	Novel plant genotype was produced
Authentication	n/a