nature portfolio

Corresponding author(s): Dr. Carlos Outeiral Rubiera
Prof. Charlotte M. Deane, MBE

Last updated by author(s): Jan 17, 2023

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	The training dataset was constructed as described in the manuscript using the sequences available at the European Nucleotide Archive on April 2022, with data code "CON", corresponding to high-quality assembled genomes. Validation datasets used to test the predictive ability of the large language model were downloaded from the original sources cited in the manuscript, and filtered to entries that could be reliably mapped to ENA-deposited cDNA sequences using UniProtKB.
Data analysis	Python packages NumPy (1.21.5), scikit-learn (1.1.2) and PyTorch (1.11.0) were used to analyse the data. Default parameters were used unless otherwise specified in the manuscript.

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

We have made available, in our website, the training set (http://opig.stats.ox.ac.uk/data/downloads/training_data.tar.gz), the heldout set (http://opig.stats.ox.ac.uk/data/downloads/heldout.tar.gz) and the weights of the trained model (http://opig.stats.ox.ac.uk/data/downloads/calm_weights.pkl).
All datasets used to test the predictive capacities of the Codon adaptation Language Model (CaLM) are available under the `data` directory on the official GitHub repository (https://github.com/oxpig/CaLM) or at the CodeOcean capsule accompanying this manuscript.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	This study did not involve human research participants.
Population characteristics	This study did not involve human research participants.
Recruitment	This study did not involve human research participants.
Ethics oversight	This study did not involve human research participants.

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	Sample size calculations were not relevant to this study.
Data exclusions	The training data, as well as the validation experiments, excluded cDNA sequences belonging to viruses, synthetic experiments, or otherwise outside of the Archaea, Eukaryota and Bacteria taxonomic classifications. Training, heldout and validation sequences were filtered to ensure that they started in a start codon, ended in a stop codon, did not have any interstitial stop codons, and did not have any unassigned nucleotides.
Replication	All experiments conducted in this manuscript were subject to cross-validation and displayed comparable results across independent folds.
Randomization	Randomization was not relevant to this study.
Blinding	Randomization was not relevant to this study.

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

Methods

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