nature portfolio

Corresponding author(s): Guo-Wei Wei

Last updated by author(s): Dec 16, 2022

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	MSA collection is from EVcouplings (v2) server: http://v2.evcouplings.org
Data analysis	Open source code: pytorch (1.10.0), scikit-learn (0.23.1), xgboost (1.5.0), scipy (1.4.1), biopython (1.79), numpy (1.19.5), h5py (2.10.0), hyperopt (0.2.2), pandas (1.1.2), pickle (4.0), tqdm (4.32.1), gudhi (3.3.0), DeepSequence (Jul 2018), theano (1.0.5), esm (0.3.1), tensorflow (1.13.0 & 2.2.0), VMD (1.9.4), scipy (1.4.1), HERMES (1.0), TAPE (Aug, 2021; https://github.com/songlab-cal/tape-neurips2019), UniRep (Mar, 2020) Non-open source package: jackal (http://honig.c2b2.columbia.edu/jackal) The custom code for data analysis is available at https://github.com/WeilabMSU/TopFit

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

There are 34 DMS datasets with experimentally measured fitness used in this work including: 32 DeepSequence datasets [8], avGFP dataset [41], and GB1 dataset [42]. The original data sources of the 32 DeepSequence datasets are provided in Supplementary Data 1 and Supplementary Note 7. Structure data were obtained from PDB database [22] and AlphaFold [38], and the specific entry ID was provided in Supplementary Data 1.
The data analyzed and generated in this work, including sequence-to-fitness datasets, op- timized structure data, multiple sequence alignments, fine-tune parameters for eUniRep mod- els, predictions from evolutionary scores for individual mutations, and sequence- and structure- based embeddings are available at https://github.com/WeilabMSU/TopFit [65] and our lab server https://weilab.math.msu.edu/Downloads/TopFit/.
Source data for Figures 3, 4, 5, and Extended Data Figures 1, 3, 5, 6, 7, 8, 9, 10 is available with this manuscript. Source data for Extended Data Figuers 2, 4 is available in Supplementary Data 2.

Human research participants

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

Reporting on sex and gender	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	The size of data size is based on the available data from literature cited.
Data exclusions	Following the convention in DeepSequence, we exclude sequences with mutations at positions that have more than 30% gaps in MSAs. In addition, we exclude wildtype sequence in evaluating data.
Replication	For each evaluation, we repeat the evaluation for 20 times with different random seeds for fixed training data size: 24, 48, 72, 96, 120, 144, 168, 192, 216, and 240, and 10 times for 80/20 train/test split and 5-fold cross validations. All replications were successful.
Randomization	For each evaluation, we randomly pick 20% of data for testing set. The remaining 80% data is used for training data selection. For 80/20 train/test split, all 80% data is used for training set. For other cases with fixed sizes (N) of training set, we randomly sample N data within these 80% data as training set for N=24, 48, ..., 240. All samples are allocated randomly.
Blinding	We are blinded to group allocation during data collection and analysis.

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