nature portfolio

Corresponding author(s): Yufeng Shen

Last updated by author(s): Apr 17, 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	ESM-2-t33_650M_UR50D (https://github.com/facebookresearch/esm) Bcftools v1.17 (https://samtools.github.io/bcftools/) LOFTEE v1.0 (https://github.com/konradjk/loftee)
Data analysis	The machine learning model is built under tensorflow v2.8.0, and statistical analysis is performed by scipy v1.8.0 and scikit-learn v1.2.2. Other custom codes for model training and data analysis could be found at Github (https://github.com/ShenLab/MisFit) and also on zenodo.

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

Data generated by this study are all accessible, including MisFit_D, MisFit_S for missense variants, and MisFit S_genefor each gene, and have been deposited in the

Zenodo (https://doi.org/10.5281/zenodo.15230898).
The data used for training in this study are available at:
gnomAD (https://gnomad.broadinstitute.org/)
zoonomia (https://zoonomiaproject.org/)
primateAI-3D (https://primateai3d.basespace.illumina.com/)
MaveDB (https://www.mavedb.org/) for deep mutational scanning data with accession numbers urn:mavedb:00000001; 00000005; 00000013; 00000035; 00000036; 00000047; 00000048; 00000049; 00000050; 00000054; 00000055; 00000057; 00000059; 00000069; 00000078; 00000095; 00000096; 00000097; 00000108.
These raw referenced data can be obtained upon application:
Allele frequency data from UK Biobank (https://www.ukbiobank.ac.uk/)
Autism data from the SPARK for autism study, including all coding variants, can be obtained from SFARI base: https://base.sfari.org
Variant prediction results for analysis are collected:
ESM-2 and ESM-1b (https://github.com/facebookresearch/esm), AlphaMissense (https://doi.org/10.5281/zenodo.8360242), gMVP (https://www.dropbox.com/s/nce1jhg3i7jw1hx/gMVP.2021-02-28.csv.gz?dl=0) from their original releases.
PrimateAI, CADD, REVEL, MPC from dbNSFP v4.3 (https://www.dbnsfp.org/).

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	Not collected.
Reporting on race, ethnicity, or other socially relevant groupings	We used European and Africa population sequencing data, which is provided by the original datasets from gnomAD (classified by genotypic data). This was based on availability.
Population characteristics	See above.
Recruitment	Not recruited by this study.
Ethics oversight	Not applicable

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	This study does not involve in new experimental design for estimating causal effect. Sequenced sample size are collected from public databases upon availability.
Data exclusions	No data were excluded from the analysis as long as the sequencing quality have passed criteria in the original database.
Replication	Codes and data are available for reproducibility.
Randomization	Data were shuffled during training.
Blinding	The datasets were pre-collected and analysis were automated and objective.

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

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.