nature portfolio

Corresponding author(s): Tommaso Mazza

Last updated by author(s): Jul 7, 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	No software was used for data collection
Data analysis	MD simulation data was obtained and analyzed using Amber 20 and AmberTools21. In-silico mutagenesis was performed with UCSF Chimera 1.16 In-silico protein systems were refined with MODELLER 9.16 Energetic analysis of protein structures performed using FoldX 5.0 Some statistical analyses were performed using SciPy v1 Most of the machine learning algorithms were borrowed from Python's scikit-learn v1.2.1. Custom scripts written with Python 3.9 Pathogenicity predictors (no version available if not provided): PolyPhen2, SIFT 5.0.3, fathmm 2.2, PROVEAN 1.3, MutationAssessor 2.0, EFIN, CADD 1.3, PANTHER, PhD-SNP, SNAP, MutationTaster2 Custom code available at https://github.com/mazzalab/playgrounds as a Colab notebook. DOI: 10.5281/zenodo.8137948

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 datasets supporting the conclusions of this article are included within the article and its Supplementary Data files.
The APOGEE 2 probabilities/classes of pathogenicity and the whole MitImpact database can be freely downloaded from http://mitimpact.css-mendel.it.
Datasets 2 and 3 are freely available from GnomAD [https://gnomad.broadinstitute.org] and HelixMTdb [https://www.helix.com/pages/mitochondrial-variant-database], respectively. Dataset 4 is freely available from MITOMAP [https://mitomap.org/MITOMAP/Benign].
ClinVar is freely available at https://ftp.ncbi.nlm.nih.gov/pub/clinvar/.

PDB IDs:
5xtc [http://doi.org/10.2210/pdb5xtc/pdb] of the Respiratory Complex I
5z62 [http://doi.org/10.2210/pdb5z62/pdb] for the Cytochrome C Oxidase
5xte [http://doi.org/10.2210/pdb5xte/pdb] for the Respiratory Complex III

Human research participants

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

Reporting on sex and gender	This research does not involve human participants
Population characteristics	See above
Recruitment	See above
Ethics oversight	See above

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 calculation was not needed for this study. Being the number of known mitochondrial pathogenic and neutral variants small, the APOGEE 2 training set included all available variants.
Data exclusions	No data was excluded
Replication	MD simulation was performed in triplicates for each investigated variant. Simulation profiles were consistent through replicas. Part of the cross-validation procedure of the machine learning workflow was repeated five times to ease the evaluation of the classifier's performance.
Randomization	Randomization was used during the training process of APOGEE, as explained in the "Machine learning workflow" section in Methods.
Blinding	No blinding process was required, being a computational tool

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