

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

Machaon's implementation and evaluation scripts are available to the community at: <https://github.com/anastasiadoulab/machaon>. The computed features on the PDB files in Spike protein structural comparisons are available at <https://zenodo.org/record/6654658>.

Data analysis

Machaon was implemented in Python v.3.8 (<https://www.python.org/>) and uses the following packages: NumPy (<https://github.com/numpy/numpy>), pandas (<https://github.com/pandas-dev/pandas>), SciPy (<https://www.scipy.org/>), scikit-learn (<https://github.com/scikit-learn/scikit-learn>), Matplotlib (<https://github.com/matplotlib/matplotlib>), seaborn (<https://github.com/mwaskom/seaborn>), BioPython (<https://github.com/biopython/biopython>), RDKit (<https://github.com/rdkit/rdkit>), Open3D (<https://github.com/isl-org/Open3D>), HDBScan (<https://github.com/scikit-learn-contrib/hdbscan>), ranky (<https://github.com/Didayolo/ranky>), DSSPParser (<https://github.com/neolei/DSSPParser>), Pebble (<https://github.com/noxdafox/pebble>), lxml (<https://github.com/lxml/lxml>), networkx (<https://github.com/networkx/networkx>), Plotly (<https://github.com/plotly/plotly.py>), UMAP (<https://github.com/lmcinnes/umap>), word_cloud (https://github.com/amueller/word_cloud), tqdm (<https://github.com/tqdm/tqdm>), beautiful soup (<https://code.launchpad.net/beautifulsoup>). A modified version of TM-align (<https://zhanglab.dcm.med.umich.edu/TM-align/>, source code retrieved at 10/2/2021) is used by the evaluation model for 3D structure similarity computation. Secondary structures are determined by DSSP (<https://github.com/PDB-REDO/dssp>).

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 utilized UniProt ID mapping resources were retrieved in 21/9/2021 at: https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping and RefSeq resources were retrieved in 6/11/2021 at: <ftp://ftp.ncbi.nlm.nih.gov/refseq/>. RCSB PDB GraphQL (<https://data.rcsb.org/graphql>), UniProt (<https://www.uniprot.org/uploadlists/>, <https://www.uniprot.org/uniprot/ACCESSION.xml/<ACCESSION-NUMBER>>) and EBI QuickGo (<https://www.ebi.ac.uk/QuickGO/services/ontology/go/terms/GO:<termid>>) online services are used as a fallback method to retrieve required data which are not present in the local static data sources.

The PDB dataset for the comparisons to Spike was retrieved in December 2020 and it was the query result for viral proteins in RCSB (<https://www.rcsb.org/docs/programmatic-access/batch-downloads-with-shell-script>). The dataset containing PDB files of the Spike protein was obtained from the same source (RCSB). Benchmark datasets were obtained from <https://github.com/rcsb/biozernike-validation> for Task 1, http://shrec2018.drugdesign.fr/shape_retrieval_shrec2018_pdb_files.tar.gz, http://shrec2018.drugdesign.fr/SHREC2018_ref.cla for Task 2 and http://ekhidna2.biocenter.helsinki.fi/dali/pdb_and_scope.tar for Task 3.

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender

Population characteristics

Recruitment

Ethics oversight

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](https://www.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

Data exclusions

Replication

Randomization

Not applicable, we are not making a comparison between groups.

Blinding

Not applicable, we are not making a comparison between groups.

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