nature research

Corresponding author(s): Daisuke Kihara

Last updated by author(s): 6/16/2022

Reporting Summary

Nature Research 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 Research 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 collecting data.
Data analysis	The DAQ program (v1.0.0) used in this study is freely available for academic use via Github, https://github.com/kiharalab/DAQ. The program is available to run on a Google Collab website, https://bit.ly/daq-score.

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 Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

The list of IDs of PDB and EMDB entries used in the datasets are provided in Supplementary Table 2, 4, 6 and 7.

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 data we used in this work is from two public databases, PDB and EMDB. The procedure of selecting entries for the datasets we used are described in Methods in details. For training, validating, and testing Emap2sec+, we used 223 EM maps and associated PDB entries. For evaluating performance of DAQ-score, we used three datasets, 35 protein chain model pairs, 399 protein structure model pairs, and 4485 protein structure models. The way these datasets are constructed are described in the Method section. We provide the procedure used also here: The first dataset of 35 protein chain model pairs (PDB2Ver dataset) was constructed as follows: First, from a total of 13,279 EMDB entries at 5 Å resolution or better, we identified those in which there exist two versions of the deposited structure in the associated wwPDB database (ftp-versioned.wwpdb.org). We then selected for analysis those in which the Cα-RMSD between the first and the revised version of the deposited structure model was 1.0 Å or larger, leaving 15 EMDB entries, which contained 35 protein chains in total. The second dataset of 399 protein structure model pairs (the PDBNR90 dataset), was constructed as follows: From cryo-EM maps of 1.5 to 5.0 A resolution, protein pairs were selected that have over 90% sequence identity with each other yet have a Cα RMSD of 1.0 Å or higher and have at least four contiguous potentially misaligned residues in their entries. The last dataset of 4485 PDB chains were collected from EM maps with resolution better than 5.0 Å. These protein models, the PDBNR1Å dataset, are non-redundant in terms of sequence and structure, having less than 90% sequence identity or more than 1.0 Å Cα RMSD with other entries.
Data exclusions	No data was excluded as long as the data match the selection condition described above and in the manuscript.
Replication	Replication is not relevant to this work because we analyzed all the protein models individually that are selected and in the datasets.
Randomization	Randomization is not relevant to this work because we analyzed all the protein models individually that are in the datasets.
Blinding	Blinding is not relevant to this work because we analyzed all the protein models individually that are in the datasets.

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
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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