nature portfolio

Corresponding author(s): Matteo Dal Peraro

Last updated by author(s): Jul 4, 2024

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

The main software and packages used are: PyTorch 1.10, Numpy 1.21.2, GEMMI 0.4.9, ColabFold 1.5.2, AlphaFold2, Amber 2016. We provide all the code to reproduce the datasets and experiments as a GitHub repository available at https://github.com/LBM-EPFL/CARBonAra.

Data analysis

The main software and packages used are: Tensorboard 2.12.2, Matplotlib 3.7.1, Blosum 2.0.2, MDTraj 1.9.6, ChimeraX 1.3. We provide all the code to reproduce the data analysis and figures as a GitHub repository available at https://github.com/LBM-EPFL/CARBonAra.

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

We provide instructions to access and reproduce the data at https://github.com/LBM-EPFL/CARBonAra.

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	n/a
Reporting on race, ethnicity, or other socially relevant groupings	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	We trained and tested the model using the largest amount of data publicly accessible. We selected the size of testing dataset to be a significant fraction of the available datasets: 70'000 subunits; ~15% of all available subunits. For all analysis of the method, we use a large enough sample size to draw significant conclusions.
Data exclusions	We excluded structures that only contains C_alpha coordinates. We established this exclusion criteria because our method uses atomic structures as input and cannot be applied to structures containing only C_alpha. Only a few structures in the full dataset are affected by this exclusion.
Replication	The results and analysis was repeated multiple times. We replicated the training for 5 models with small variations of the architecture of the neural network. The training of the 5 independent models were successful: the analysis shows comparable performance between all 5 models. We picked the best model to showcase our method.
Randomization	We ensure that the set of structures are independent by clustering structures with more than 30% sequence identity together and fold similarity using C.A.T.H classification. The training, testing and validation set were randomly sampled from the set independent clusters. The structures were randomly sampled during training. We randomly sampled independent structures for validation and benchmarking.
Blinding	n/a: the data is publicly available and the analysis was automated so there was no human intervention that could introduce bias.

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	n/a
Novel plant genotypes	n/a
Authentication	n/a