nature research

Corresponding author(s): Bonnie Berger
Alex Noble

Last updated by author(s): August 1, 2019

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 Authors & Referees 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	Leginon (Suloway, 2005) was used for cryoEM single particle data collection.
Data analysis	Topaz was developed for particle picking as described in this work. Topaz is available on GitHub at https://github.com/tbepler/topaz. Structure determination and 2d class averaging was perfomed with cryoSPARC v0.6.5. For additional data processing and analysis, software used was MotionCor2 v1.2.1, Appion v3.3beta, CTFFIND4, UCSF Chimera v1.13.1, 3DFSC, DoG Picker 2, FindEM v2, crYOLO v1.1.3, DeepPicker, and EMAN2.22.

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/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

Single particle half maps, full sharpened maps, and masks for T20S proteasome, 80S ribosome, rabbit muscle aldolase, and the Toll receptor (DoG, template, and Topaz picks) have been deposited to the Electron Microscopy Data Bank (EMDB) with accession codes EMD-9194, EMD-9201, EMD-9202, EMD-9206, EMD-9207, EMD-9208, EMD-9209, EMD-9210, EMD-9211, EMD-20529, EMD-20531, and EMD-20532. The full rabbit muscle aldolase dataset has been deposited to the Electron Microscopy Pilot Image Archive (EMPIAR) with accession code EMPIAR-10215.

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	When comparing PU learning methods, we randomly sampled training particles and fit models ten times. No sample-size calculations were performed. This number was chosen to get the best mean and variance estimates within a reasonable runtime.
Data exclusions	No data were excluded.
Replication	Particle picking with topaz and reconstruction was replicated across three cryoEM datasets. As discussed in the manuscript, performance of Topaz was well reproduced across all three datasets. We also reproduced the PU learning comparison on two cryoEM datasets where we confirmed for both that the generalize-expectation criteria approaches outperformed the non-negative risk estimator and naive baselines.
Randomization	Training particles were sampled randomly.
Blinding	Blinding was not relevant. No data was collected and analysis was performed using well established metrics without investigator intervention.

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
☒	☐	Animals and other organisms
☒	☐	Human research participants
☒	☐	Clinical data

Methods

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