nature research

Corresponding author(s): Lucy Jane Colwell

Last updated by author(s): Nov 22, 2021

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	No software was used to collect data for this manuscript, all data was obtained from the Pfam database v32.0 or v34.0. The data splits built in this manuscript have been made freely available from https://console.cloud.google.com/storage/browser/brain-genomics-public/research/proteins/pfam/random_split and https://console.cloud.google.com/storage/browser/brain-genomics-public/research/proteins/pfam/clustered_split.
Data analysis	The Top pick HMM model was implemented using HMMER 3.1b2. The TensorFlow API, specifically tensorflow-gpu v1.15.4, was used to implement and train all deep models using the architectures described in Methods. Code that documents model training using python v3.7 is available on github at \url{https://github.com/google-research/google-research/tree/master/using_dl_to_annotate_protein_universe}. The training and validation datasets used for creating each model are available as described in the preceding section. Trained models are available in Google Cloud Storage at \url{https://console.cloud.google.com/storage/browser/brain-genomics-public/research/proteins/pfam/models/single_domain_per_sequence_zipped_models}, including the ensembles trained on the Pfam seed random split, Pfam seed clustered split, Pfam full random split (all Pfam v32.0), and the models used to generate Pfam-N v34.0. ProtCNN inference was run using a custom python script that (a) read in FASTA records and (b) ran inference of the ProtCNN as a TensorFlow SavedModel. An interactive notebook that demonstrates inference using ProtCNN is available at \url{https://colab.research.google.com/github/google-research/google-research/blob/master/using_dl_to_annotate_protein_universe/neural_network/Neural_network_accuracy_on_random_seed_split.ipynb}. An interactive notebook showing use of the trained models to produce Pfam class predictions as well as embeddings is available in github at \url{https://colab.sandbox.google.com/github/google-research/google-research/blob/master/using_dl_to_annotate_protein_universe/Using_Deep_Learning_to_Annotate_the_Protein_Universe.ipynb}

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

The data splits described in this manuscript are available for download at \url{https://console.cloud.google.com/storage/browser/brain-genomics-public/research/proteins/pfam/random_split}, \url{https://console.cloud.google.com/storage/browser/brain-genomics-public/research/proteins/pfam/clustered_split} and an interactive notebook for data loading is available at \url{https://www.kaggle.com/googleai/pfam-seed-random-split}. Model predictions are freely available to download as part of the Pfam v34.0 release from \url{http://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam34.0/}.

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 sample size was determined by the size of the Pfam data bases.
Data exclusions	We used the Pfam seed and full data sets. No data was excluded from the analysis.
Replication	59 replicate ProtCNN models were trained using different random initialization of parameters.
Randomization	Sequences were split (i) randomly and (ii) randomly by distance into train, tune and held-out test sets.
Blinding	All models were blinded to the labels of all sequences in all held-out test sets.

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