

## 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

The source code for collecting and preprocessing all the data used in this study is available for research and non-commercial use at: <https://github.com/flatironinstitute/DeepFRI>  
We used the following tools and collecting and preprocessing our data:  
cd-hit-v4.8.1-2019-0228  
DeepGO-v1.0.0  
CATH FunFam-v4.2  
FFPred3  
BLAST 2.7.1+

#### Data analysis

The source code for training the DeepFRI model, together with neural network weights are available for research and non-commercial use at: <https://github.com/flatironinstitute/DeepFRI>  
A web service of our method is available at: <https://beta.deepfri.flatironinstitute.org/>

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

Our training, validation and test data splits are available from:  
<https://github.com/flatironinstitute/DeepFRI>

They are obtained from the following publicly available databases:

BioLiP: <https://zhanglab.ccmb.med.umich.edu/BioLiP/>

PDB: <https://www.rcsb.org/>

SIFTS: <https://www.ebi.ac.uk/pdbe/docs/sifts/>

Pfam: <http://pfam.xfam.org/>

CATH: <https://www.cathdb.info/>

SWISS-MODEL: <https://swissmodel.expasy.org/>

UniProt: <https://www.uniprot.org/>

CSA: <http://www.ebi.ac.uk/thornton-srv/databases/CSA/>

## 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	All the training/test/validation data sets used in the training of our model are collected from publicly available databases and their sizes are described in the Methods section in the main manuscript. The sizes of the training/test/validation sets are chosen such that each protein included in these sets is annotated with at least one function and each function used in our analysis has at least 50 training samples.
Data exclusions	PDB chains with no annotations and with > 1200 and < 60 residues are excluded from our analysis.
Replication	All attempts at replication were successful.
Randomization	This is not relevant to our study.
Blinding	This is not relevant to our study.

## 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
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging