UNIVERSITY OF WARSAW

MIM

FACULTY
OF MATHEMATICS, INFORMATICS
AND MECHANICS

UNIWERSYTET WARSZAWSKI

# Architecture of large projects in bioinformatics (ADP)

*Lecture 08*

Łukasz P. Kozłowski

Warsaw, 2025

lukaskoz@mimuw.edu.pl

# Tackling big biological problems

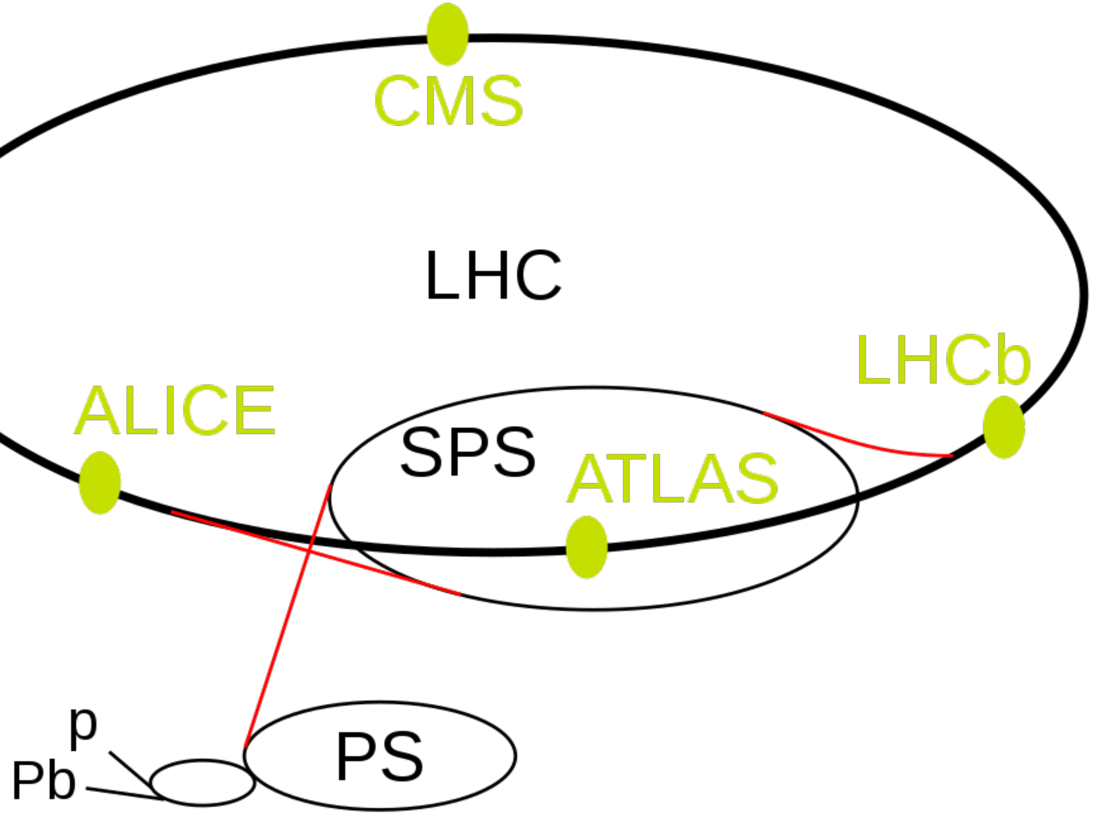
## Extra big consortia/initiatives

## Consortia

## Scientific Competitions

# ATLAS experiment

## ATLAS (A Toroidal LHC ApparatuS)

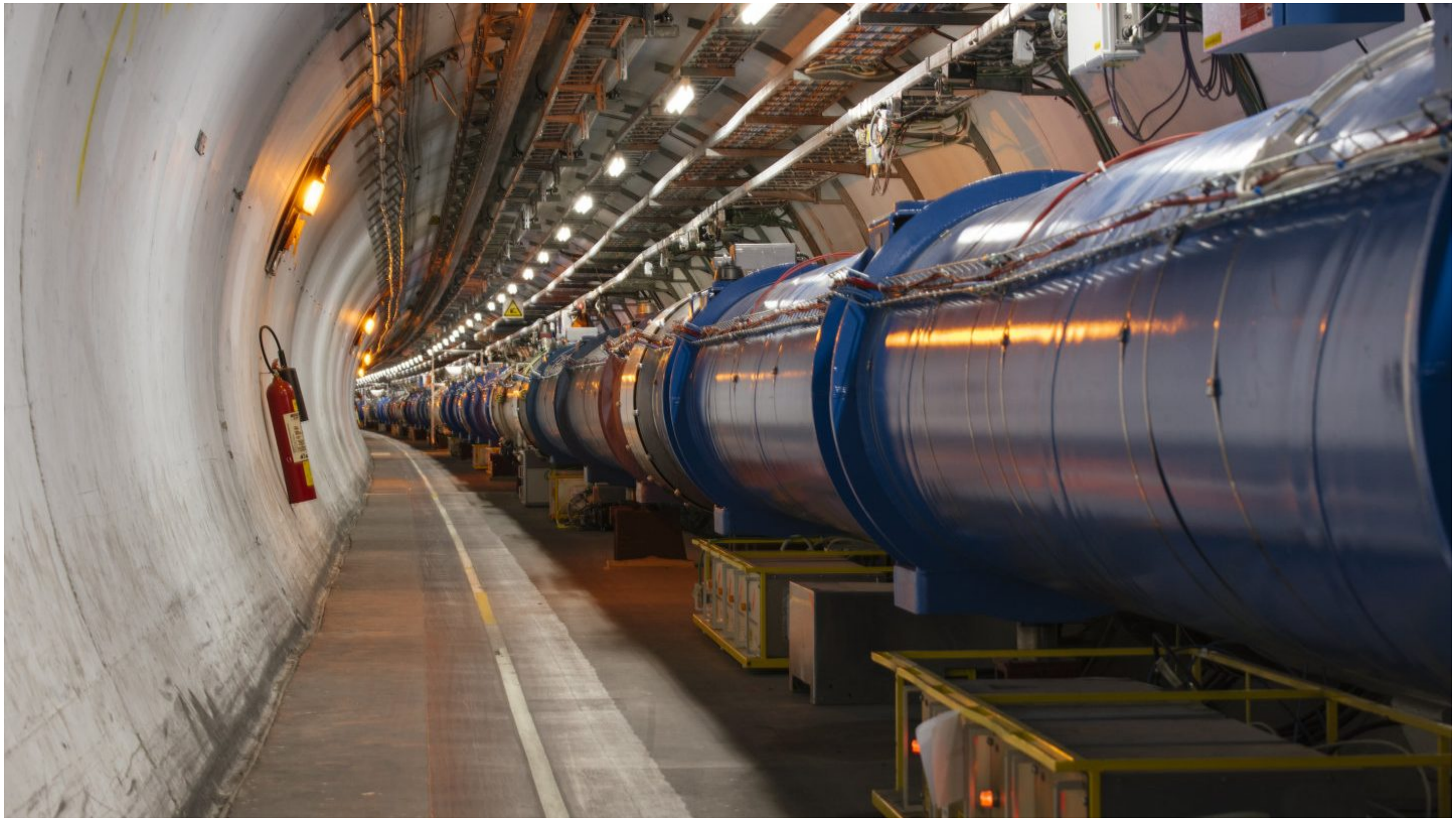


# ATLAS experiment

## ATLAS (A Toroidal LHC ApparatuS)

# ATLAS experiment

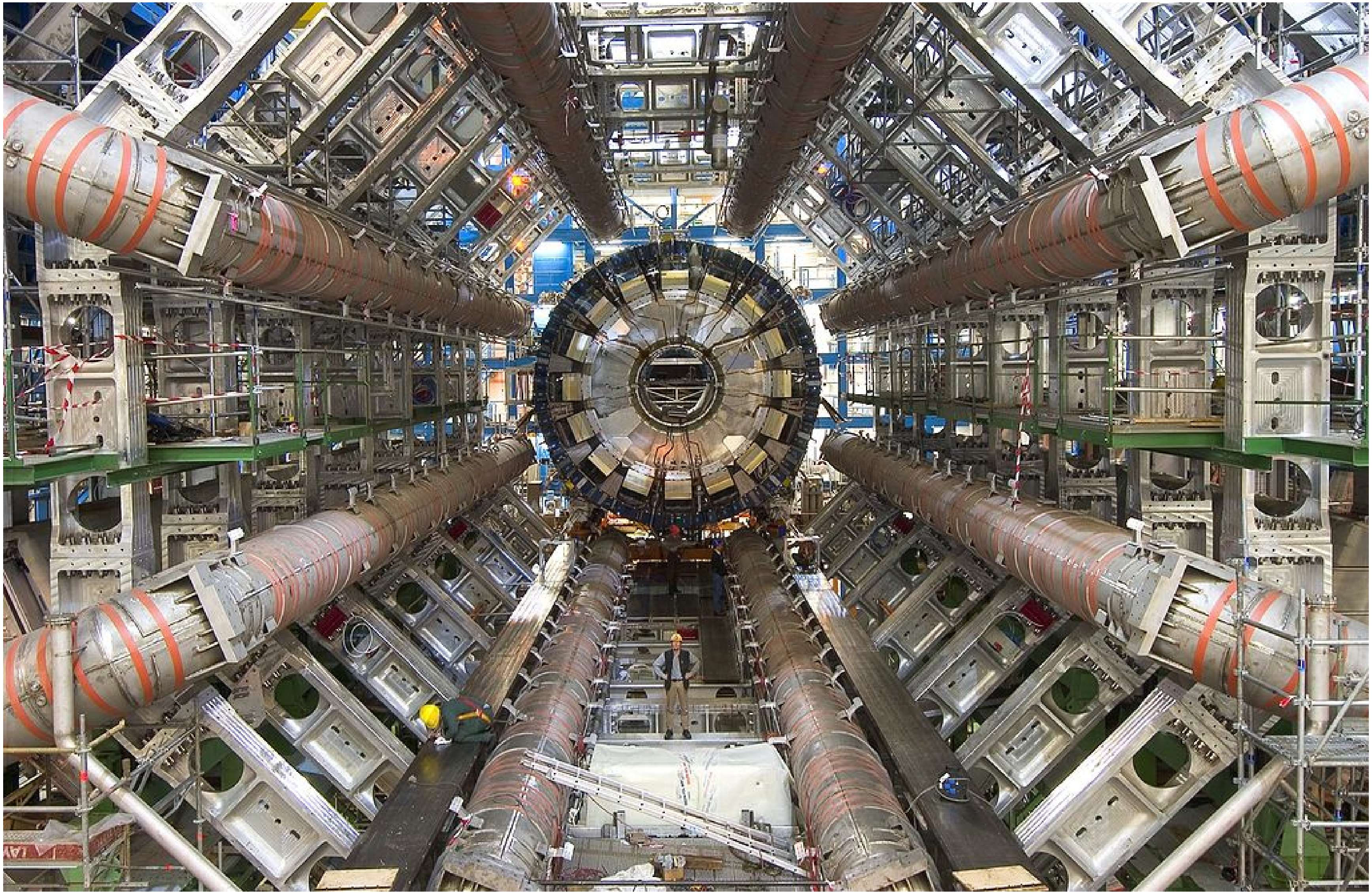
**ATLAS (A Toroidal LHC ApparatuS)**



# ATLAS experiment

## ATLAS (A Toroidal LHC ApparatuS)

# ATLAS experiment



ATLAS detector is 46 metres long, 25 metres in diameter, and weighs about 7,000 tonnes; it contains some 3000 km of cable

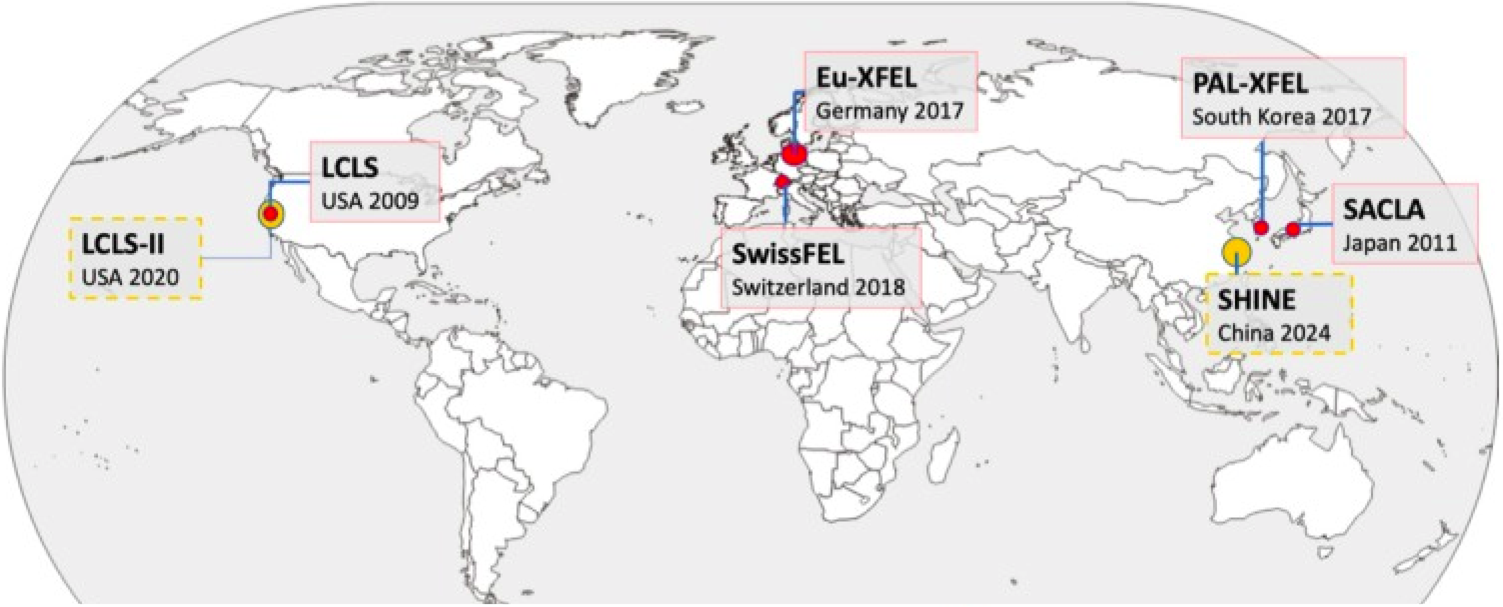
# ATLAS experiment

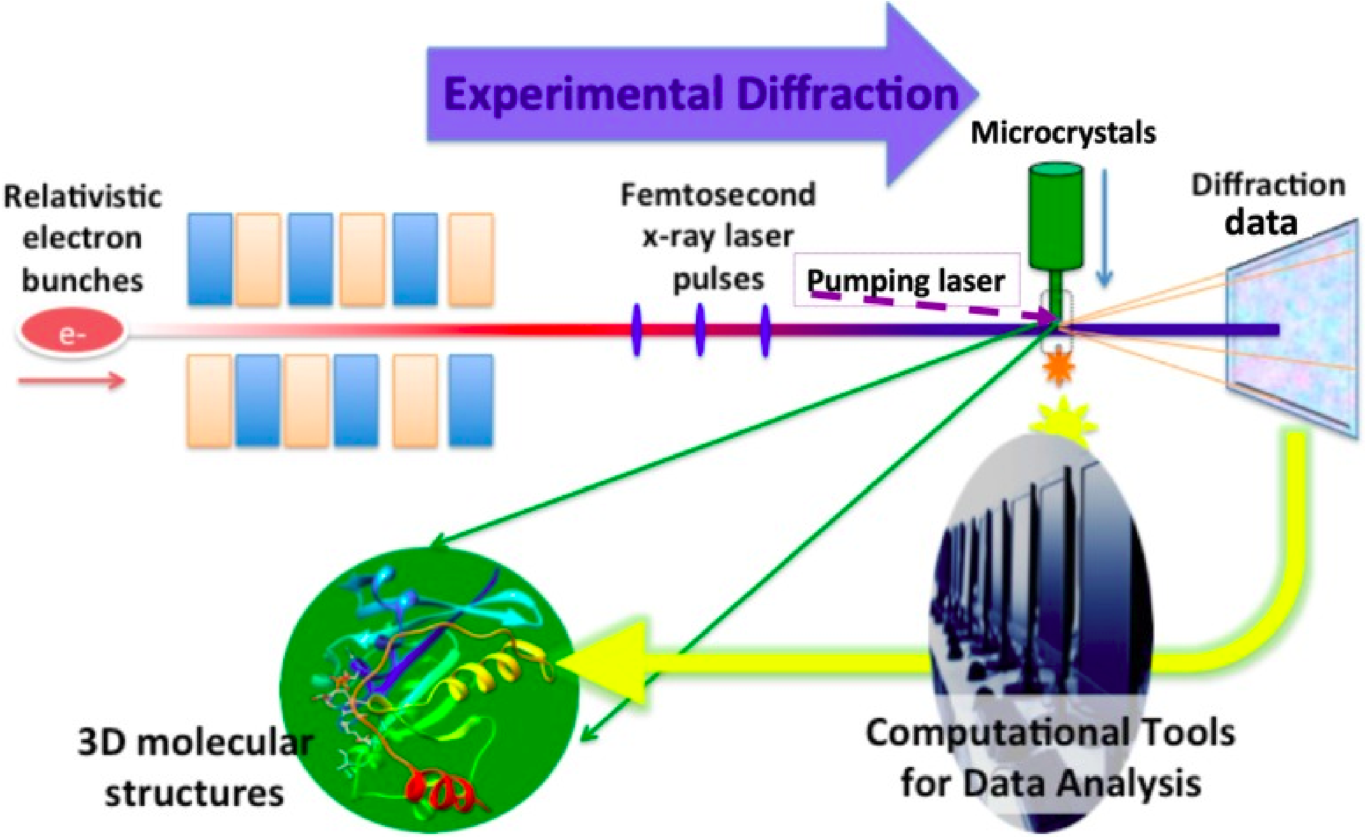
The experiment is a collaboration involving roughly 10,000 physicists from hundreds institutions in >100 countries

Budget of €7.5 billion

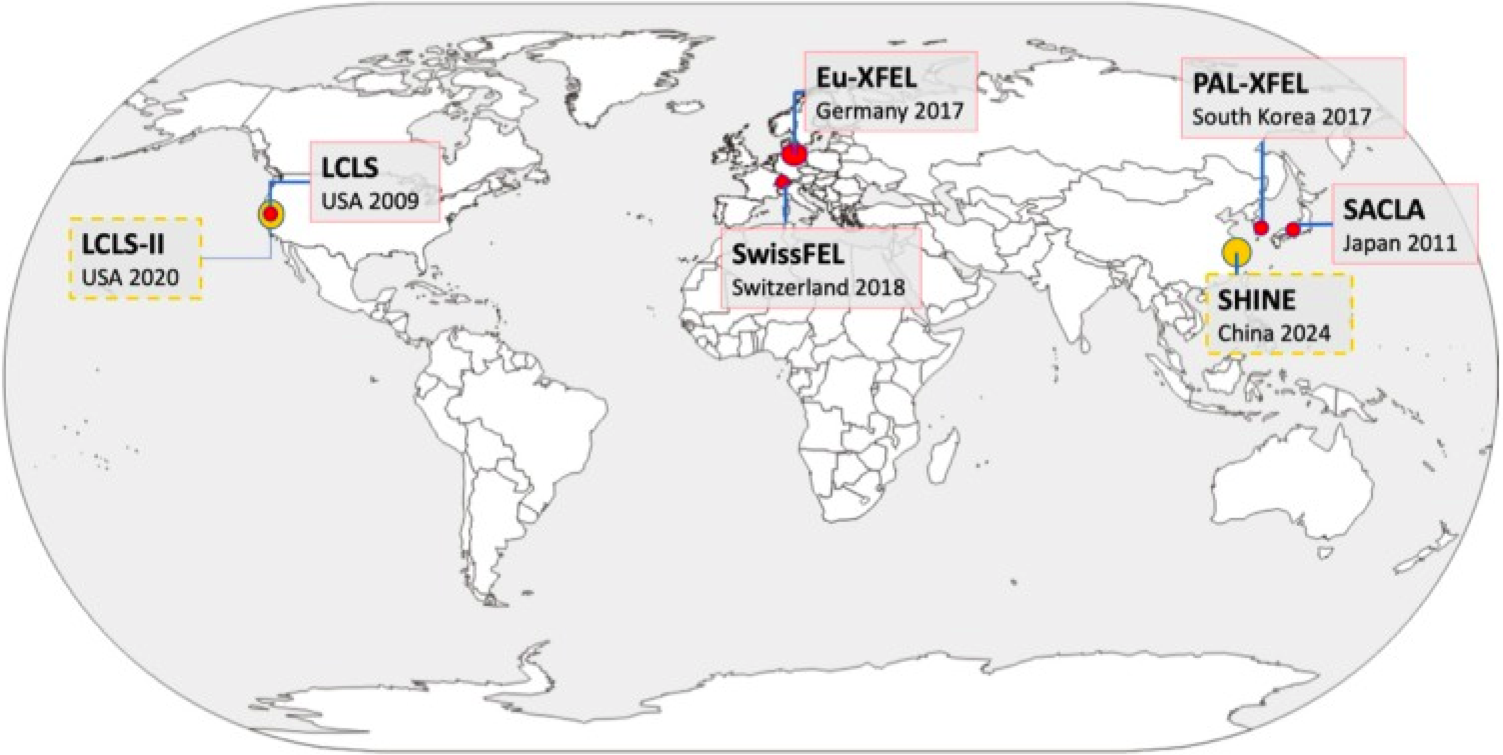
First ring build in 1971–1984

# Extra big consortia/initiatives




Experimental Diffraction
Microcrystals
Relativistic electron bunches
Femtosecond x-ray laser pulses
Diffraction data
Pumping laser
e-
3D molecular structures
Computational Tools for Data Analysis

# Extra big consortia/initiatives



European XFEL

3.4-kilometre (2.1 mi) long tunnel

Cost for the construction and commissioning of the facility is as of 2017 estimated at €1.22 billion

# Human Brain Project

Future Emerging Technologies (FET) Flagships from EU

121 partners from universities, research institutes and companies in 20 countries

Launched in 2013

Budget of €1 billion (until 2019)

**Brain, graphene and quantum technologies**

https://www.humanbrainproject.eu

# Human Brain Project

Future Emerging Technologies (FET) Flagships from EU

121 partners from universities, research institutes and companies in 20 countries

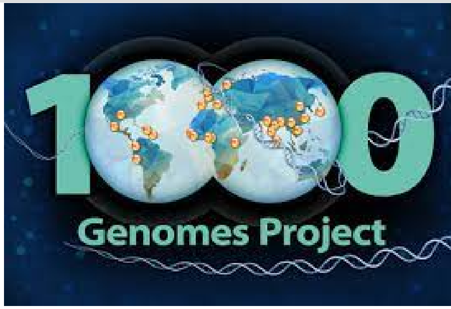
Launched in 2013

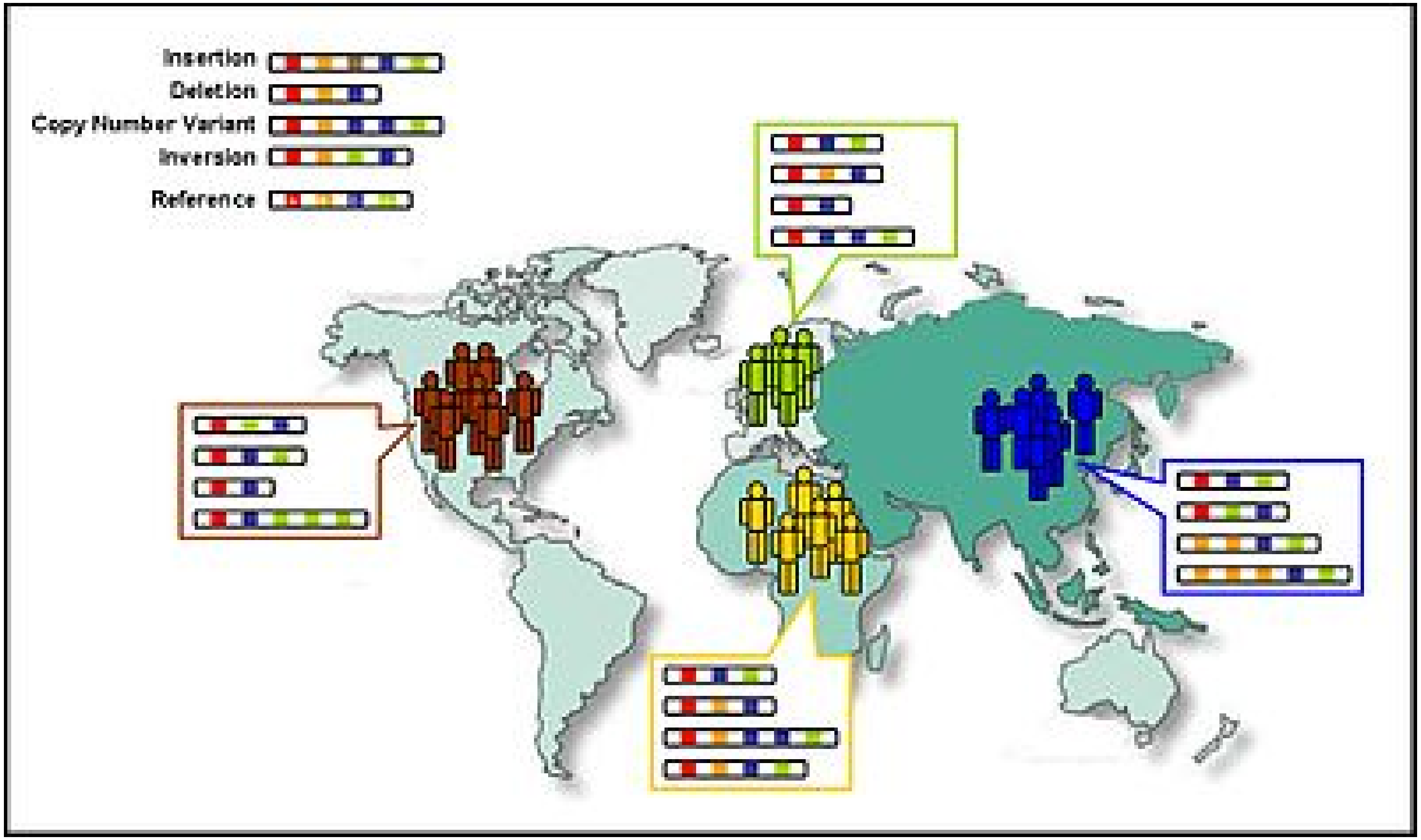
Budget of €1 billion (until 2019)

**Brain, graphene and quantum technologies**

16 of these projects will collaborate with the **Graphene** Flagship and nine with the **Human Brain Project**. These projects will be funded by a total budget of € 16.4 million and are expected to start between December 2019 and March 2020

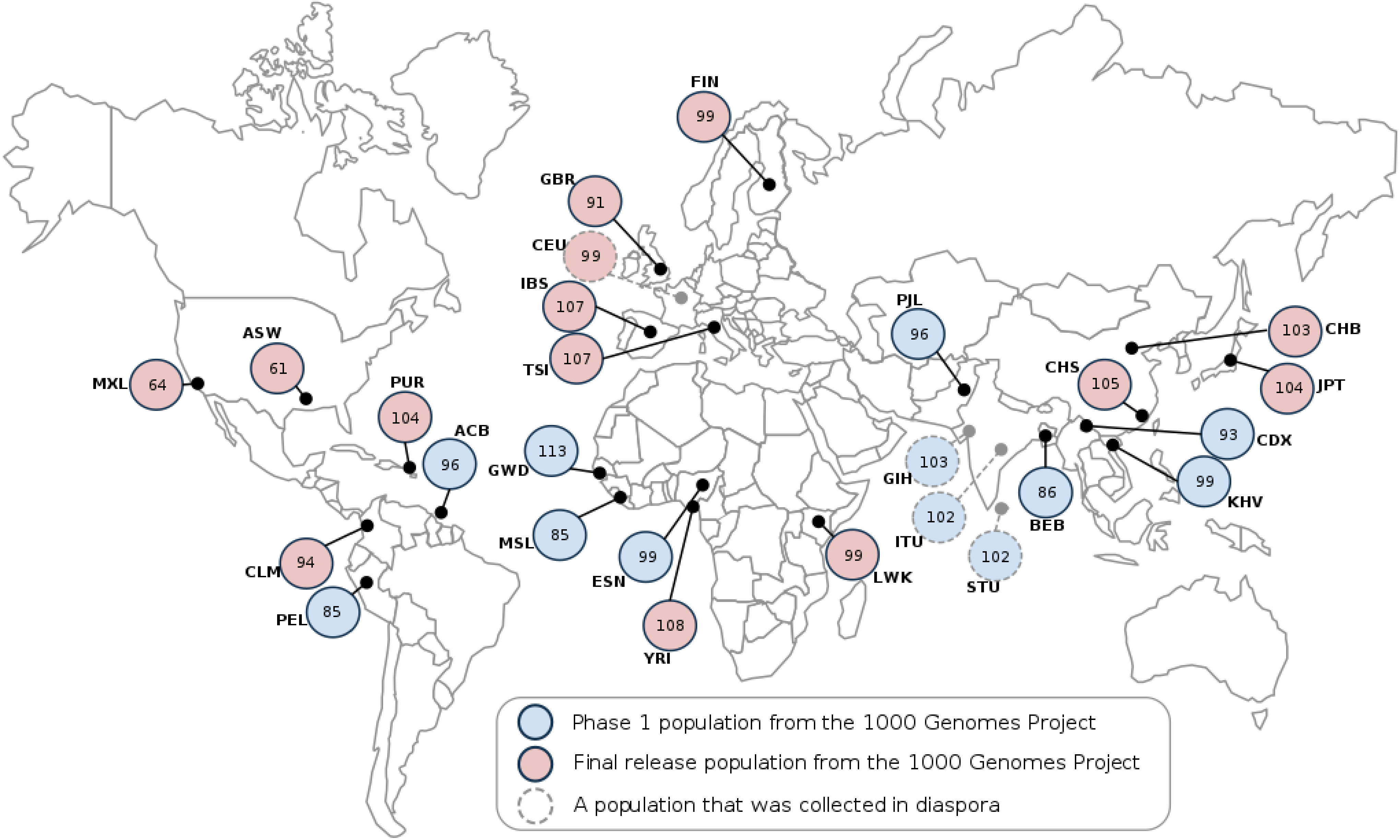
# 1000 Genomes Project





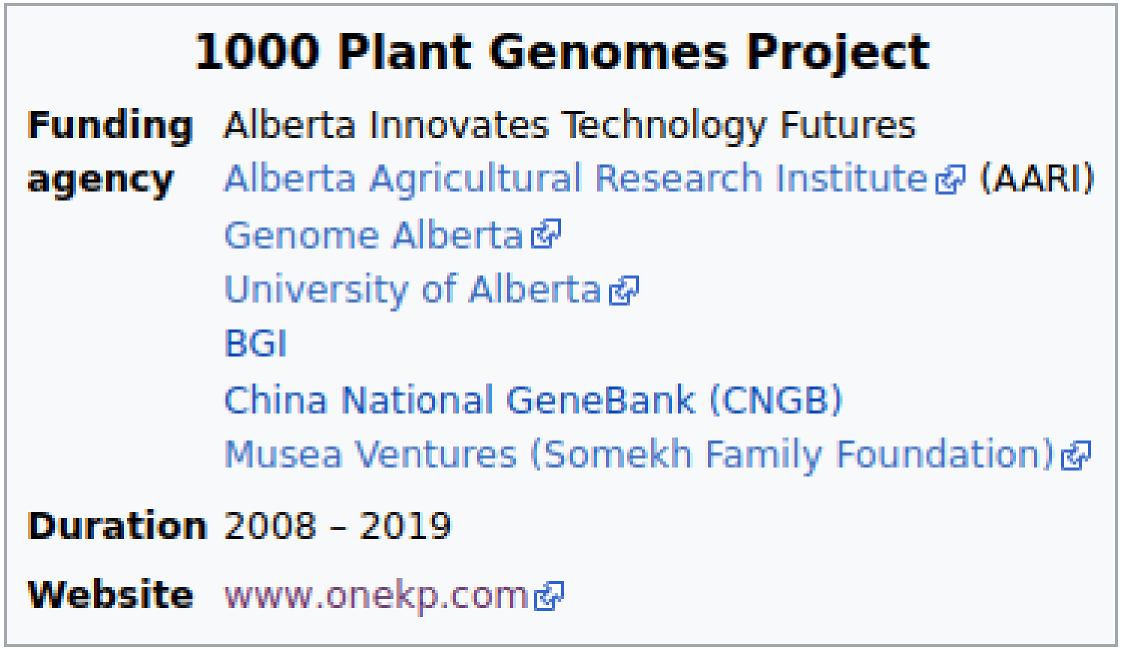
https://www.internationalgenome.org

# 1000 Genomes Project



# 1000 Plant Genomes Project (1KP)

## Followed by 10,000 Plant Genome Project



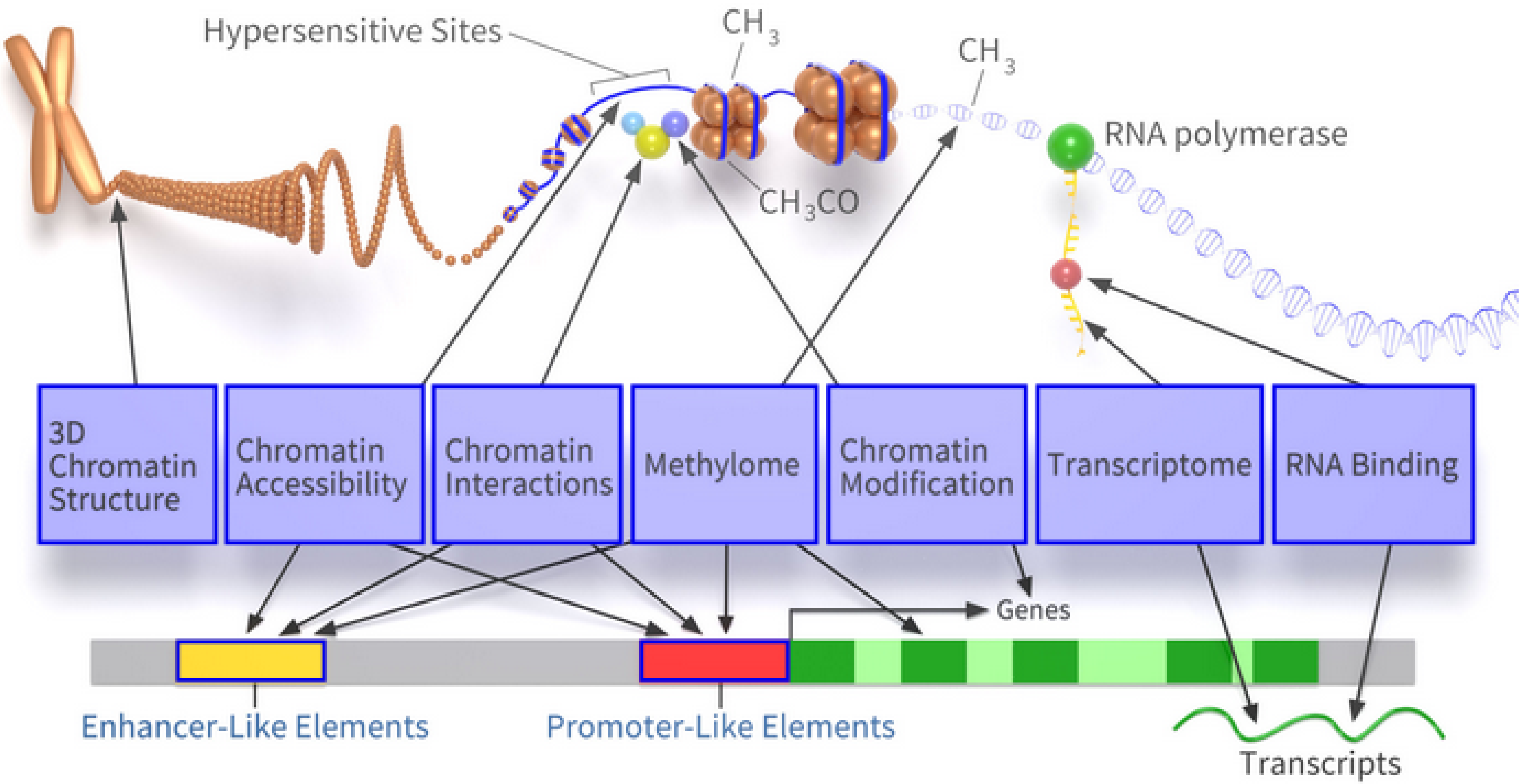
**1000 Plant Genomes Project**

| | |
|---|---|
| **Funding agency** | Alberta Innovates Technology Futures<br>Alberta Agricultural Research Institute (AARI)<br>Genome Alberta<br>University of Alberta<br>BGI<br>China National GeneBank (CNGB)<br>Musea Ventures (Somekh Family Foundation) |
| **Duration** | 2008 – 2019 |
| **Website** | www.onekp.com |

The international journal of science / 31 October 2019
nature
outline Prostate cancer
ROOTS OF DIVERSITY
Transcriptome analysis illuminates evolution of the world's green plants
A history of ethics The long and bumpy road to responsible research
Ancient climate A snapshot of $CO_2$ in the atmosphere more than 1 million years ago
Insects in decline Ten-year survey offers strong evidence of falling numbers

**https://sites.google.com/a/ualberta.ca/onekp/**

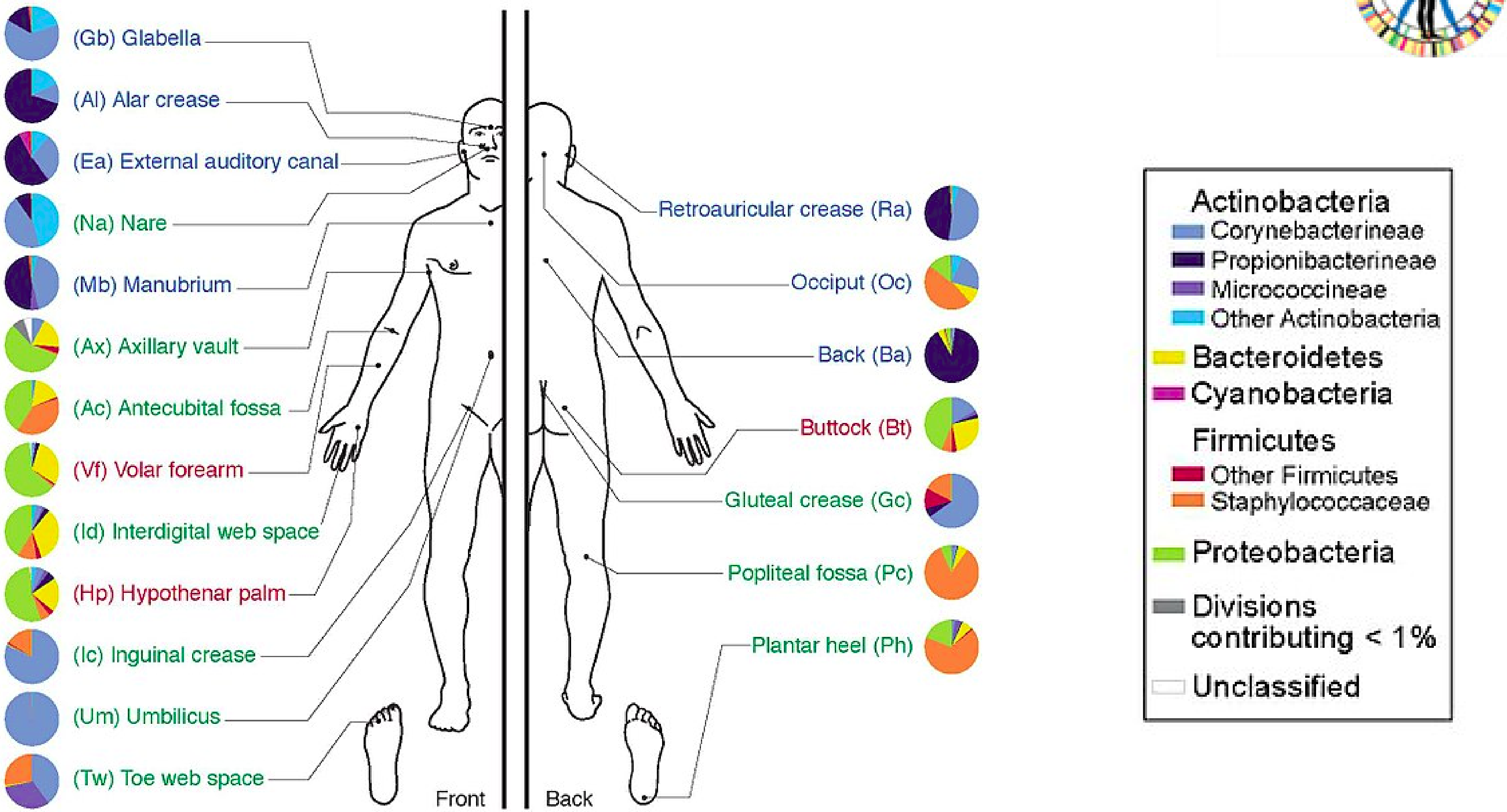
**Encyclopedia of DNA Elements (ENCODE) is a public research project which aims to identify functional elements in the human genome**



**http://encodeproject.org/**

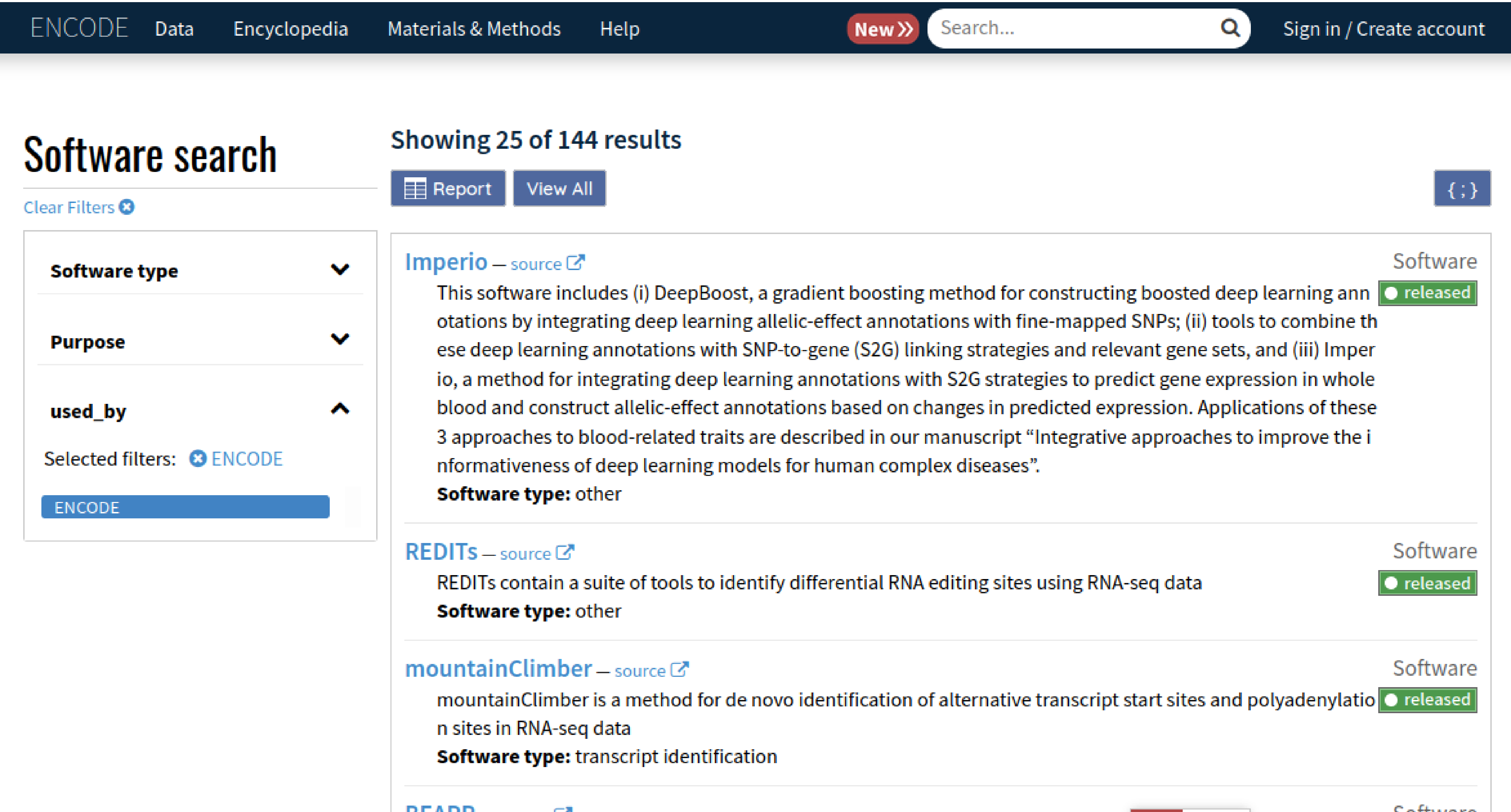
# Extra big consortia/initiatives

## Human Microbiome Project



**https://hmpdacc.org**

# Human Microbiome Project



## Software search

Selected filters: ENCODE

Showing 25 of 144 results

**Imperio** — source

Software

released

This software includes (i) DeepBoost, a gradient boosting method for constructing boosted deep learning ann otations by integrating deep learning allelic-effect annotations with fine-mapped SNPs; (ii) tools to combine th ese deep learning annotations with SNP-to-gene (S2G) linking strategies and relevant gene sets, and (iii) Imper io, a method for integrating deep learning annotations with S2G strategies to predict gene expression in whole blood and construct allelic-effect annotations based on changes in predicted expression. Applications of these 3 approaches to blood-related traits are described in our manuscript “Integrative approaches to improve the i nformativeness of deep learning models for human complex diseases”.

**Software type:** other

**REDITs** — source

Software

released

REDITs contain a suite of tools to identify differential RNA editing sites using RNA-seq data

**Software type:** other

**mountainClimber** — source

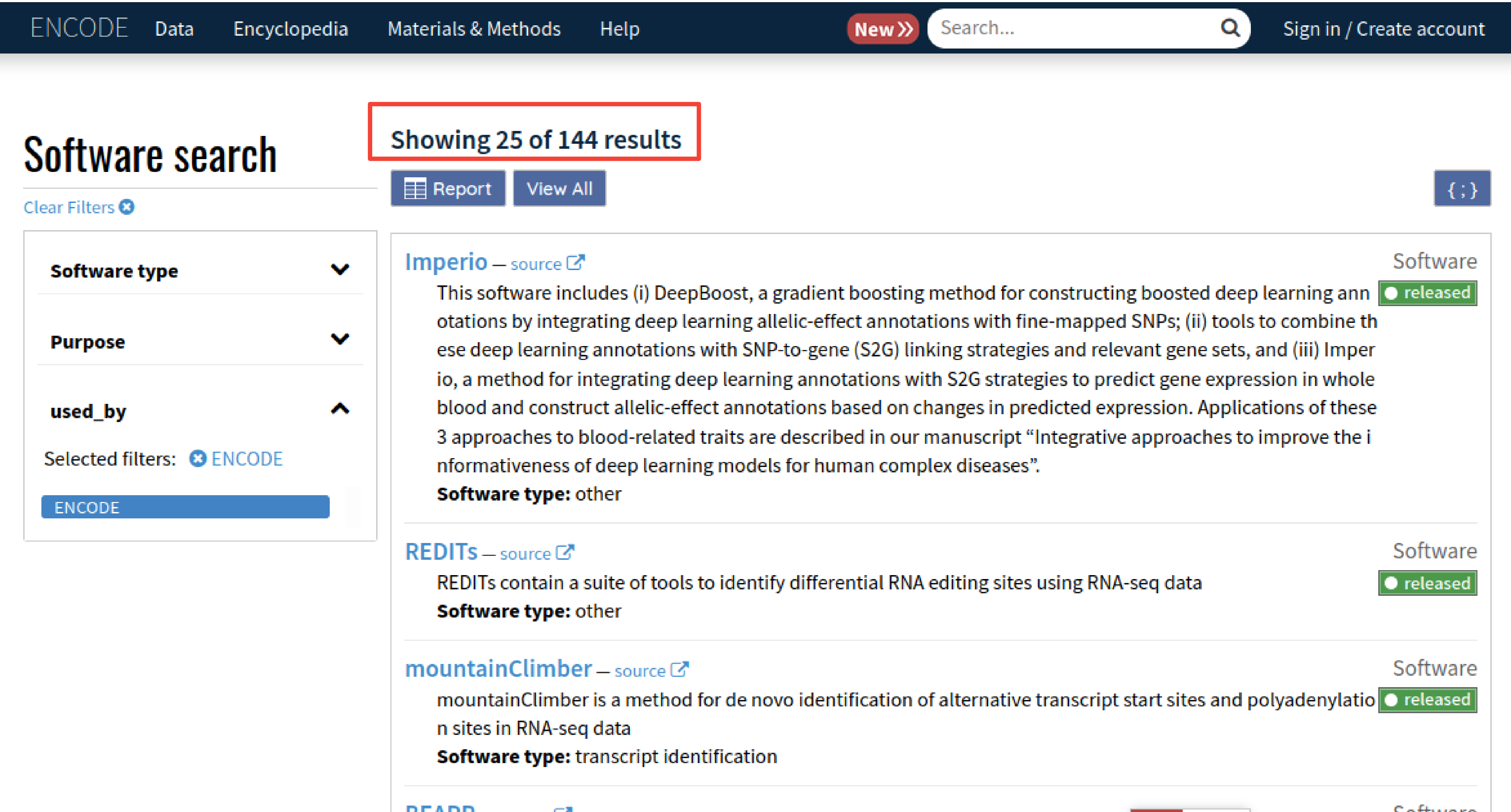
Software

released

mountainClimber is a method for de novo identification of alternative transcript start sites and polyadenylatio n sites in RNA-seq data

**Software type:** transcript identification

# Human Microbiome Project



## Software search

Clear Filters

Software type

Purpose

used_by

Selected filters: ENCODE

ENCODE

Showing 25 of 144 results

**Imperio** — source

Software

released

This software includes (i) DeepBoost, a gradient boosting method for constructing boosted deep learning annotations by integrating deep learning allelic-effect annotations with fine-mapped SNPs; (ii) tools to combine these deep learning annotations with SNP-to-gene (S2G) linking strategies and relevant gene sets, and (iii) Imperio, a method for integrating deep learning annotations with S2G strategies to predict gene expression in whole blood and construct allelic-effect annotations based on changes in predicted expression. Applications of these 3 approaches to blood-related traits are described in our manuscript "Integrative approaches to improve the informativeness of deep learning models for human complex diseases".

**Software type:** other

**REDITs** — source

Software

released

REDITs contain a suite of tools to identify differential RNA editing sites using RNA-seq data

**Software type:** other

**mountainClimber** — source

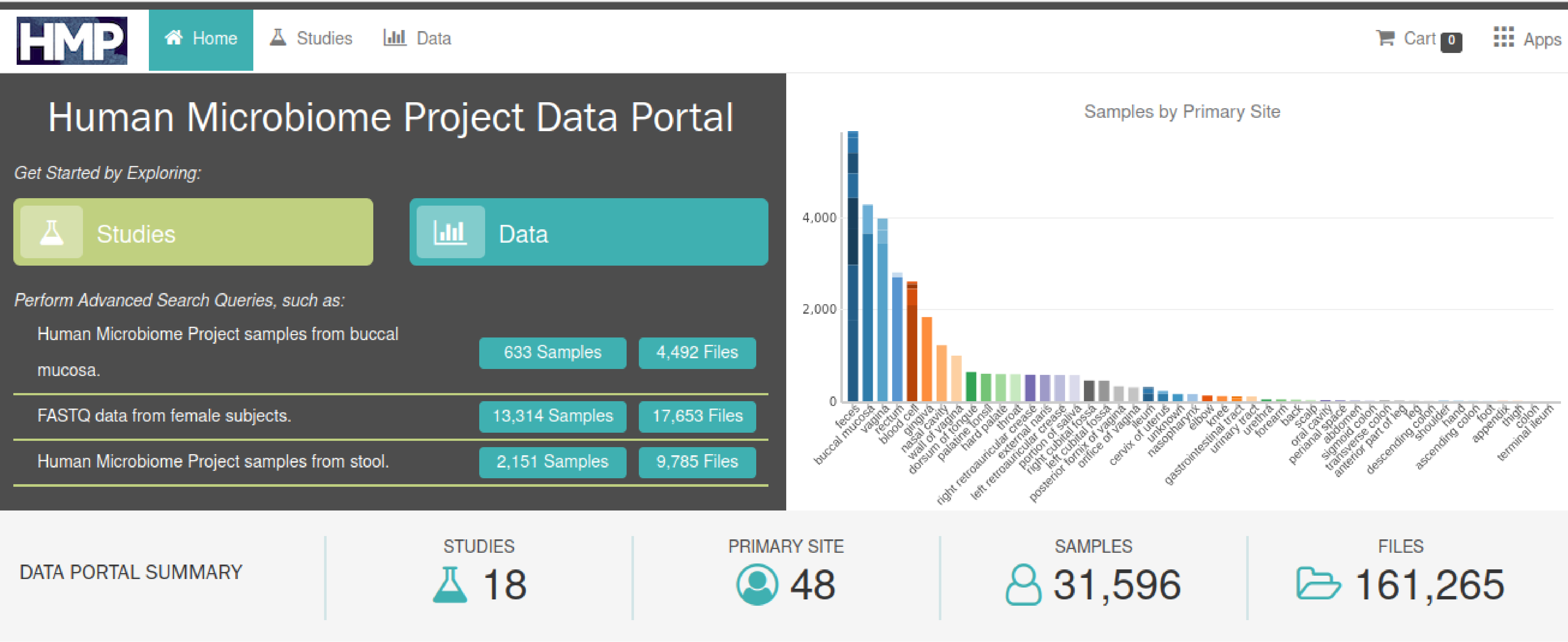
Software

released

mountainClimber is a method for de novo identification of alternative transcript start sites and polyadenylation sites in RNA-seq data

**Software type:** transcript identification

# Human Microbiome Project



HMP Home Studies Data Cart 0 Apps

## Human Microbiome Project Data Portal

*Get Started by Exploring:*

Studies Data

*Perform Advanced Search Queries, such as:*

| | | |
|---|---|---|
| Human Microbiome Project samples from buccal mucosa. | 633 Samples | 4,492 Files |
| FASTQ data from female subjects. | 13,314 Samples | 17,653 Files |
| Human Microbiome Project samples from stool. | 2,151 Samples | 9,785 Files |

| DATA PORTAL SUMMARY | STUDIES | PRIMARY SITE | SAMPLES | FILES |
|---|---|---|---|---|
| | 18 | 48 | 31,596 | 161,265 |

**https://hmpdacc.org**

biobank uk

Enabling scientific discoveries that improve human health

https://www.ukbiobank.ac.uk/

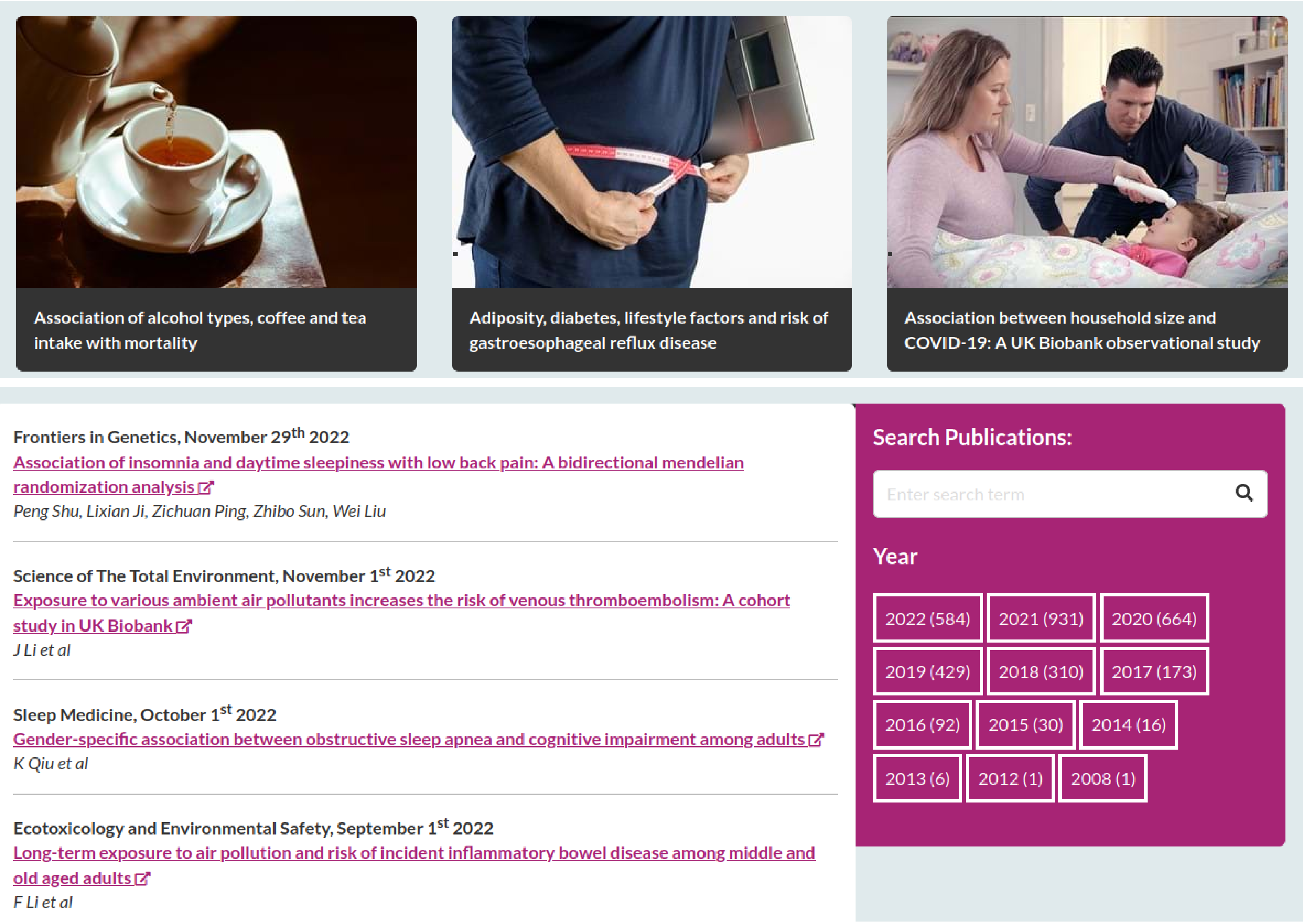
UK Biobank is a large-scale biomedical database and research resource, containing in-depth genetic and health information from **half a million UK participants**

**- NGS data**
**- Magnetic Resonance Imaging (MRI) data from the brain, heart and abdomen (>60k)**
**...**

Association of alcohol types, coffee and tea intake with mortality

Adiposity, diabetes, lifestyle factors and risk of gastroesophageal reflux disease

Association between household size and COVID-19: A UK Biobank observational study



Association of alcohol types, coffee and tea intake with mortality

Adiposity, diabetes, lifestyle factors and risk of gastroesophageal reflux disease

Association between household size and COVID-19: A UK Biobank observational study

Frontiers in Genetics, November 29th 2022

Association of insomnia and daytime sleepiness with low back pain: A bidirectional mendelian randomization analysis

*Peng Shu, Lixian Ji, Zichuan Ping, Zhibo Sun, Wei Liu*

Science of The Total Environment, November 1st 2022

Exposure to various ambient air pollutants increases the risk of venous thromboembolism: A cohort study in UK Biobank

*J Li et al*

Sleep Medicine, October 1st 2022

Gender-specific association between obstructive sleep apnea and cognitive impairment among adults

*K Qiu et al*

Ecotoxicology and Environmental Safety, September 1st 2022

Long-term exposure to air pollution and risk of incident inflammatory bowel disease among middle and old aged adults

*F Li et al*

biobank uk

Enabling scientific discoveries that improve human health

https://www.ukbiobank.ac.uk/

| Description | Tier 1 | Tier 2 | Tier 3 |
|---|---|---|---|
| Core data<br>• Questionnaires and physical measurements<br>• Health Outcome phenotypes<br>• Linked health data<br>• Web-based questionnaires | ✓ | ✓ | ✓ |
| Assay data and enhanced measures<br>• Biochemical and haematological assays<br>• Other platform based assays<br>• Measured and imputed genotypes<br>• Other enhancements |  | ✓ | ✓ |
| Very large datasets<br>• Imaging data *<br>• Other large-scale assay data<br>• Whole genome sequence data<br>• Whole exome sequence data |  |  | ✓<br>Via platform only |
| First 3 years - access to data with scheduled updates | £3,000 | £6,000<br>(+£3,000 vs Tier 1) | £9,000<br>(+£3,000 vs Tier 2) |
| Additional Institution fee - each additional institution added to an application | £1,000 for first 3 years (£500 p.a. extension) |  |  |
| Low & Middle Income Countries and Student Researchers ** - access to all datasets via the Research Analysis Platform (full fees apply to downloaded data) | £500 for first 3 years (£175 p.a. extension) |  |  |

# Scientific Competitions

# Scientific Competitions

Critical Assessment of Techniques for Protein Structure Prediction (CASP)

Collective experiment for blind RNA structure prediction (RNA-Puzzles)

Critical Assessment of Prediction of Interactions (CAPRI)

Critical Assessment of Functional Annotation (CAFA)

Critical Assessment of Microarray Data Analysis (CAMDA)

Genome Annotation Assessment Project (GASP)

Bone X-Ray Deep Learning Competition

LUng Nodule Analysis 2016

# Critical Assessment of Techniques for Protein Structure Prediction (CASP)

**CASP is a community-wide, worldwide experiment for protein structure prediction taking place every two years since 1994**

Met-Glu-Leu-Gly-Leu-Gly-Gly-Leu-Ser-Thr-Leu-Ser-His-Cys-Pro
Trp-Pro-Arg-Gln-Gln-Pro-Ala-Leu-Trp-Pro-Thr-Leu-Ala-Ala-Leu
Ala-Leu-Leu-Ser-Ser-Val-Ala-Glu-Ala-Ser-Leu-Gly-Ser-Ala-Pro
Arg-Ser-Pro-Ala-Pro-Arg-Glu-Gly-Pro-Pro-Pro-Val-Leu-Ala-Ser
Pro-Ala-Gly-His-Leu-Pro-Gly-Gly-Arg-Thr-Ala-Arg-Trp-Cys-Ser
Gly-Arg-Ala-Arg-Arg-Pro-Pro-Pro-Gln-Pro-Ser-Arg-Pro-Ala-Pro
Pro-Pro-Pro-Ala-Pro-Pro-Ser-Ala-Leu-Pro-Arg-Gly-Gly-Arg-Ala
Ala-Arg-Ala-Gly-Gly-Pro-Gly-Ser-Arg-Ala-Arg-Ala-Ala-Gly-Ala
Arg-Gly-Cys-Arg-Leu-Arg-Ser-Gln-Leu-Val-Pro-Val-Arg-Ala-Leu
Gly-Leu-Gly-His-Arg-Ser-Asp-Glu-Leu-Val-Arg-Phe-Arg-Phe-Cys
Ser-Gly-Ser-Cys-Arg-Arg-Ala-Arg-Ser-Pro-His-Asp-Leu-Ser-Leu
Ala-Ser-Leu-Leu-Gly-Ala-Gly-Ala-Leu-Arg-Pro-Pro-Pro-Gly-Ser
Arg-Pro-Val-Ser-Gln-Pro-Cys-Cys-Arg-Pro-Thr-Arg-Tyr-Glu-Ala
Val-Ser-Phe-Met-Asp-Val-Asn-Ser-Thr-Trp-Arg-Thr-Val-Asp-Arg
Leu-Ser-Ala-Thr-Ala-Cys-Gly-Cys-Leu-Gly

Anfinsen

# Scientific Competitions

## Critical Assessment of Techniques for Protein Structure Prediction (CASP)

**CASP is a community-wide, worldwide experiment for protein structure prediction taking place every two years since 1994**

**Every second spring-summer around 100 targets* are released**

**Targets – protein sequences for which the structure has been solved recently (not Available publicly e.g. not in PDB)**

## Blind benchmark

## Critical Assessment of Techniques for Protein Structure Prediction (CASP)

**CASP is a community-wide, worldwide experiment for protein structure prediction taking place every two years since 1994**

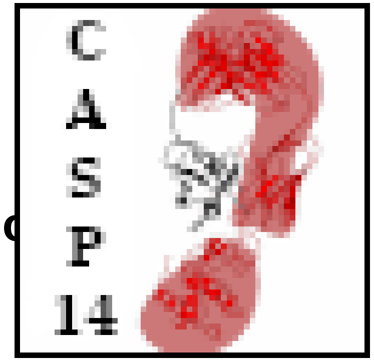
**Every second spring-summer around 100 targets* are released**

**Targets – protein sequences for which the structure has been solved recently (not Available publicly e.g. not in PDB)**

# Blind benchmark

**Categories: servers (72 h) and humans (3 weeks)**
**homology modeling & Free Modeling**

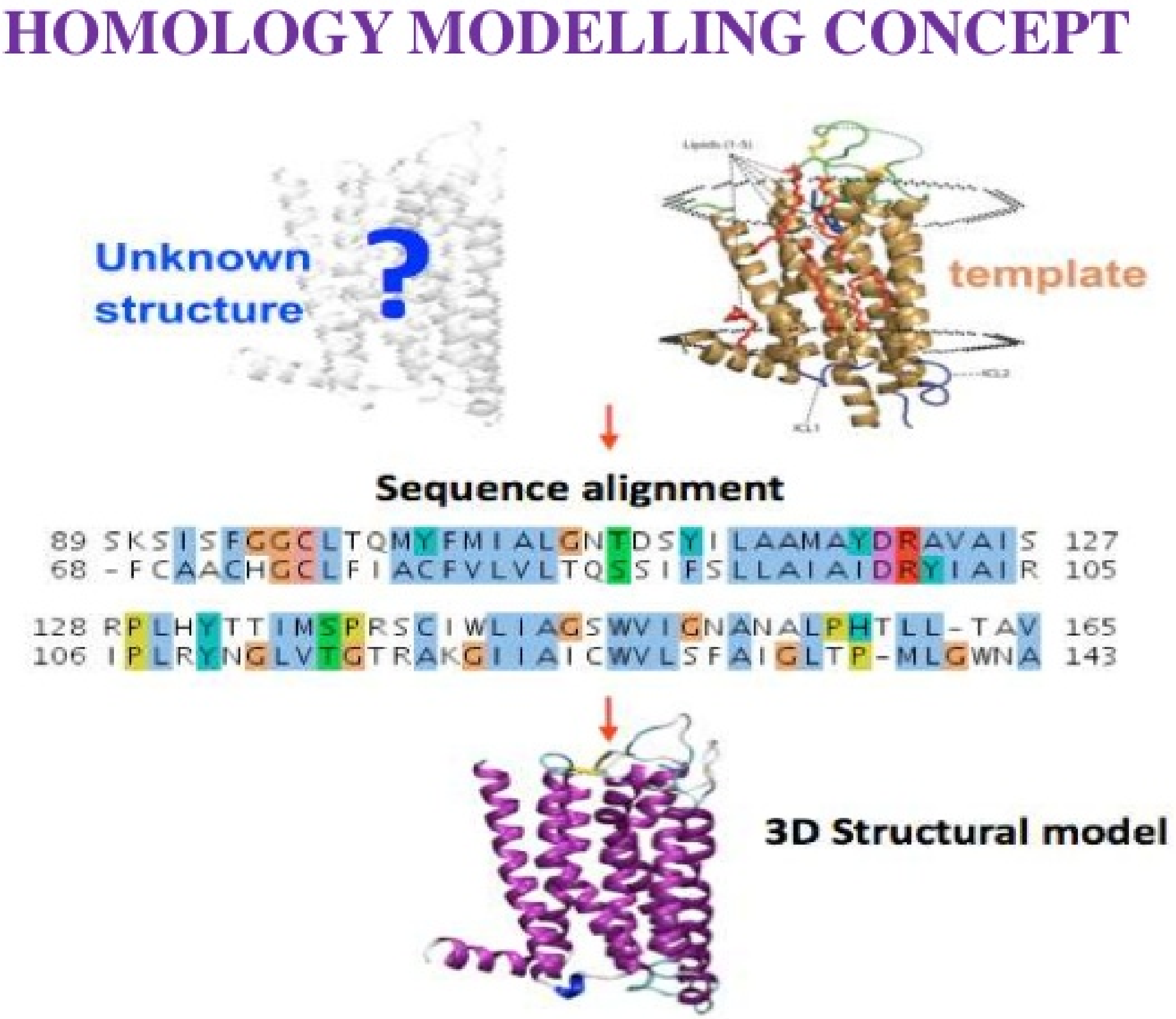
# Critical Assessment of Techniques for Protein Structure Prediction (CASP)



Evaluation of the results is carried out in the following prediction categ

- tertiary structure prediction (all CASPs)
- secondary structure prediction (dropped after CASP5)
- prediction of structure complexes (CASP2 only;
  a separate experiment - CAPRI—carries on this subject)
- residue-residue contact prediction (starting CASP4)
- disordered regions prediction (starting CASP5)
- domain boundary prediction (CASP6–CASP8)
- function prediction (starting CASP6)
- model quality assessment (starting CASP7)
- model refinement (starting CASP7)
- high-accuracy template-based prediction (starting CASP7)

## Tertiary structure prediction (all CASPs)

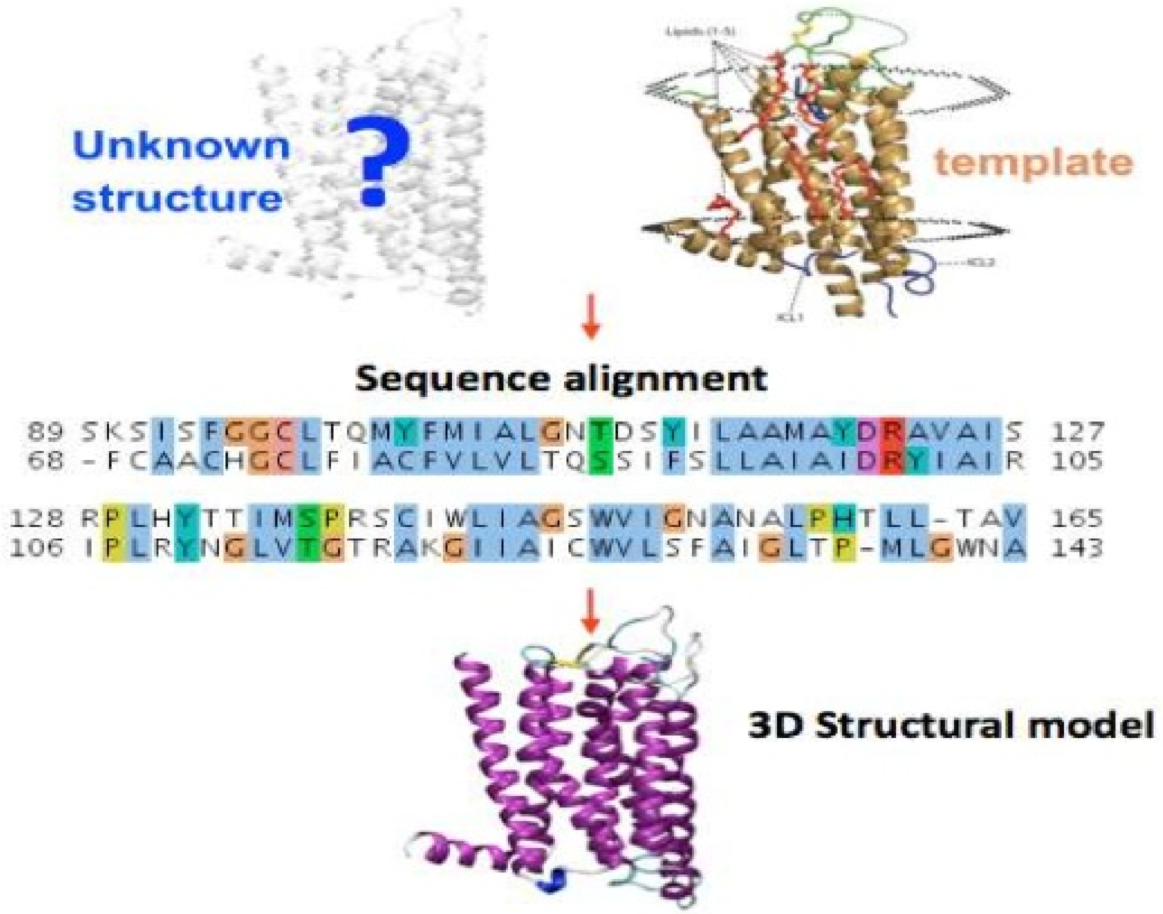


```
89  SKSISFGGCLTQMYFMIALGNTDSYILAAMAYDRAVAIS  127
68  -FCAACHGCLFIACFVLVLTQSSIFSLLAIAIDRYIAIR  105

128 RPLHYTTIMSPRSCIWLIAGSWVIGNANALPHTLL-TAV  165
106 IPLRYNGLVTGTRAKGIIAICWVLSFAIGLTP-MLGWNA  143
```

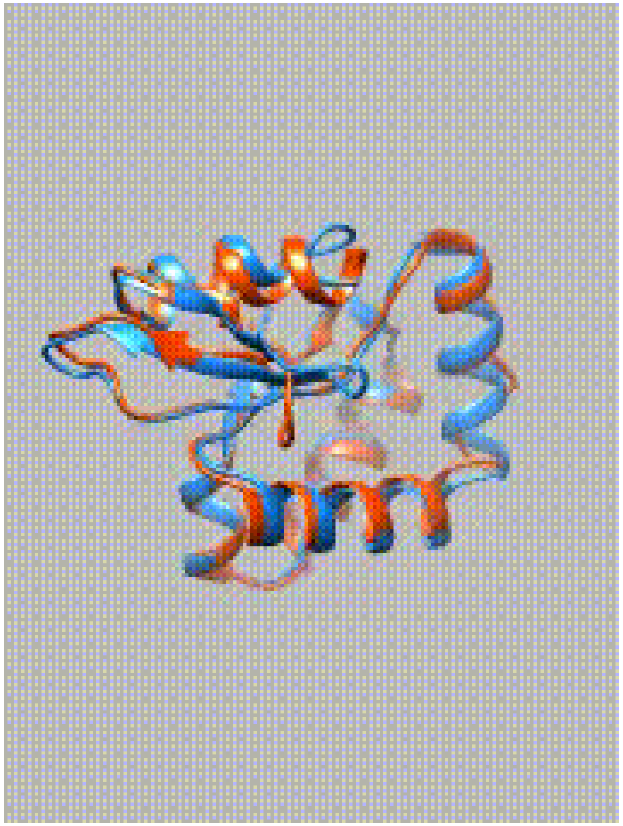
# Tertiary structure prediction (all CASPs)

CASP 14

## HOMOLOGY MODELLING CONCEPT

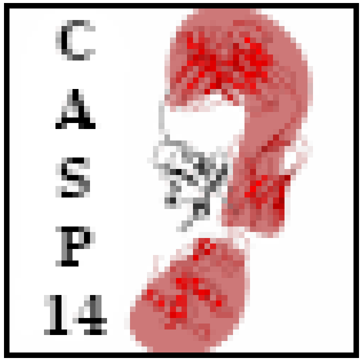


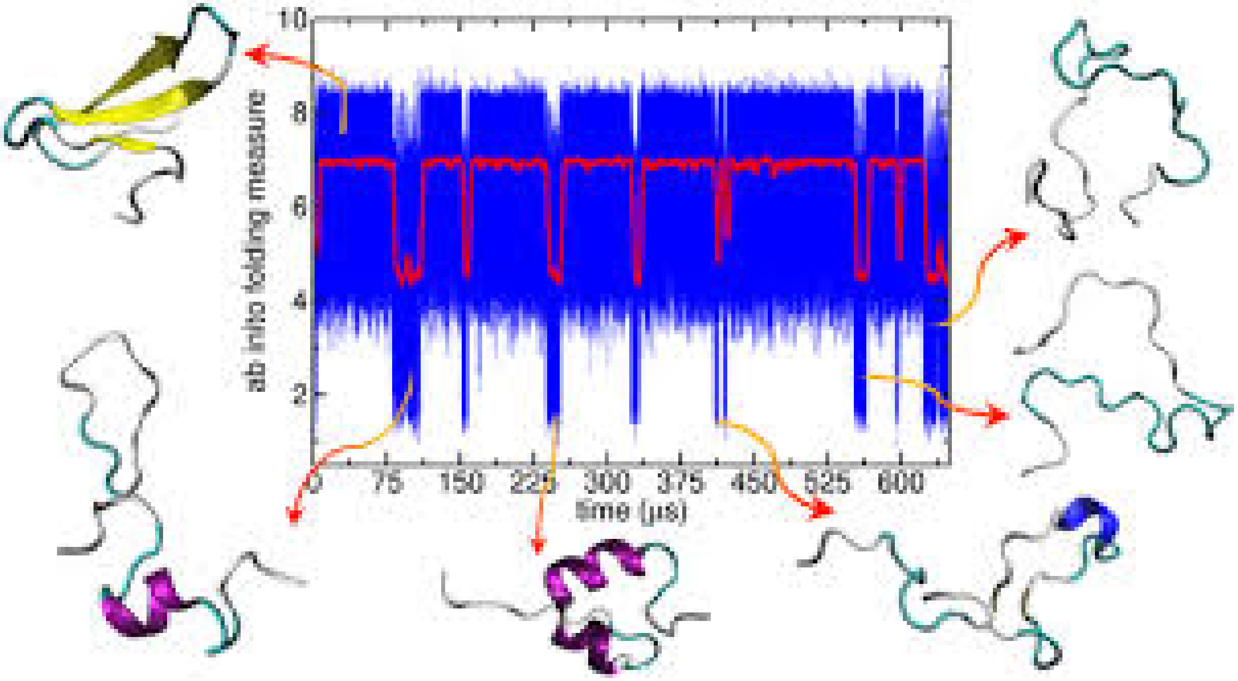
```
89  SKSISFGGCLTQMYFMIALGNTDSYILAAMAYDRAVAIS  127
68  -FCAACHGCLFIACFVLVLTQSSIFSLLAIAIDRYIAIR  105

128 RPLHYTTIMSPRSCIWLIAGSWVIGNANALPHTLL-TAV  165
106 IPLRYNGLVTGTRAKGIIAICWVLSFAIGLTP-MLGWNA  143
```



target T0868-D1 (orange)
model 330_2 (blue): GDT_TS=87
best template: 2cw6 (seq.id= 4.2%)

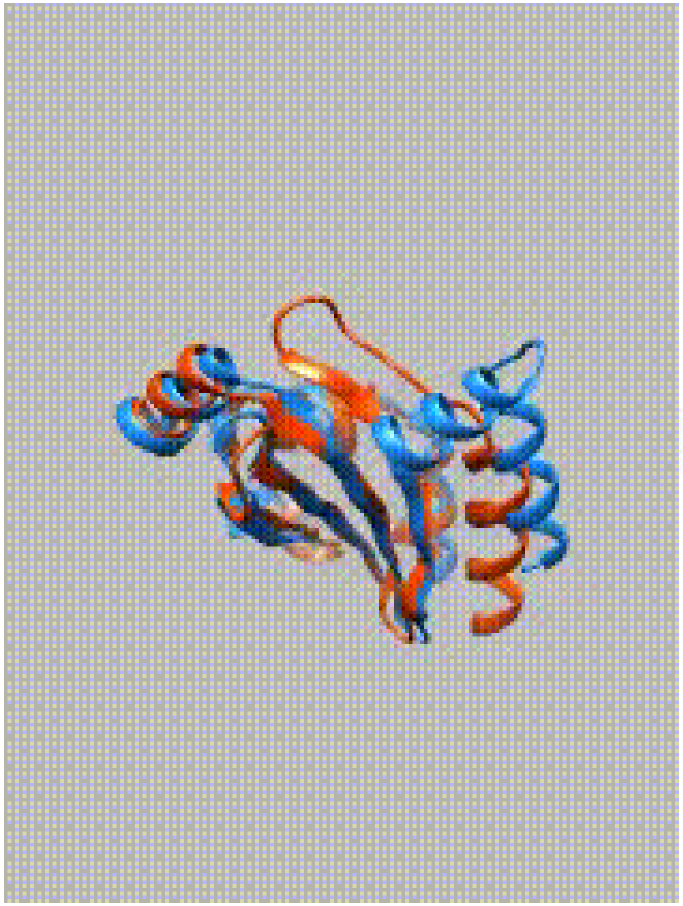
# Tertiary structure prediction (all CASPs)



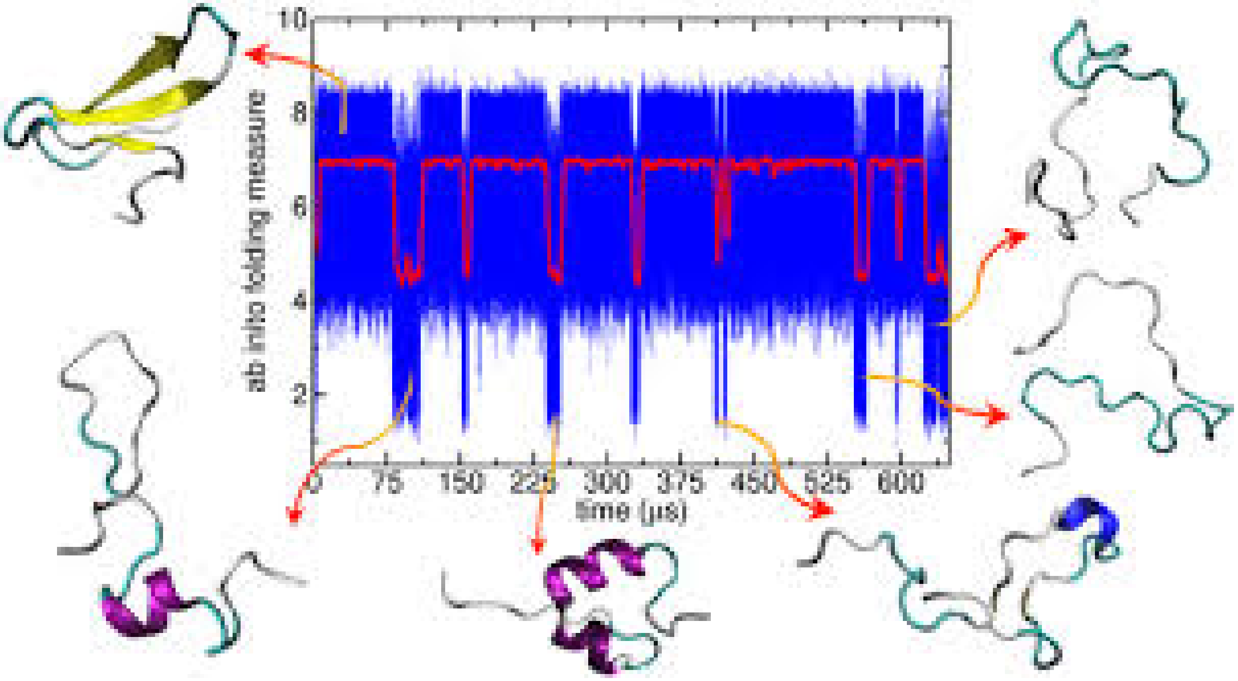


## Tertiary structure prediction (all CASPs)

CASP 14



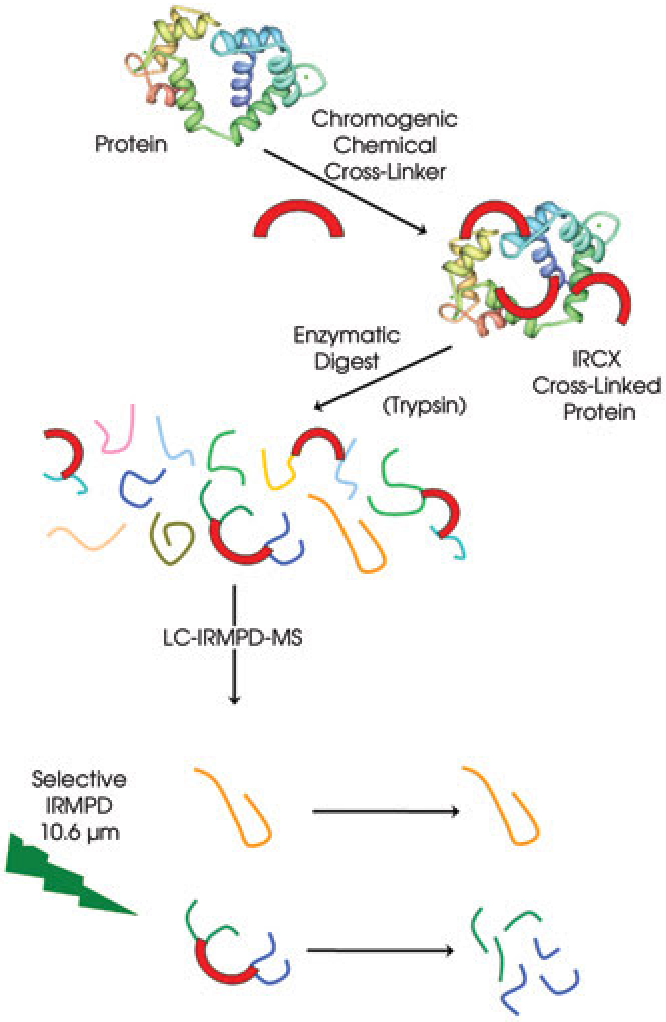
CASP9: T0581-D1
5 model 170_1: GDT_TS=71

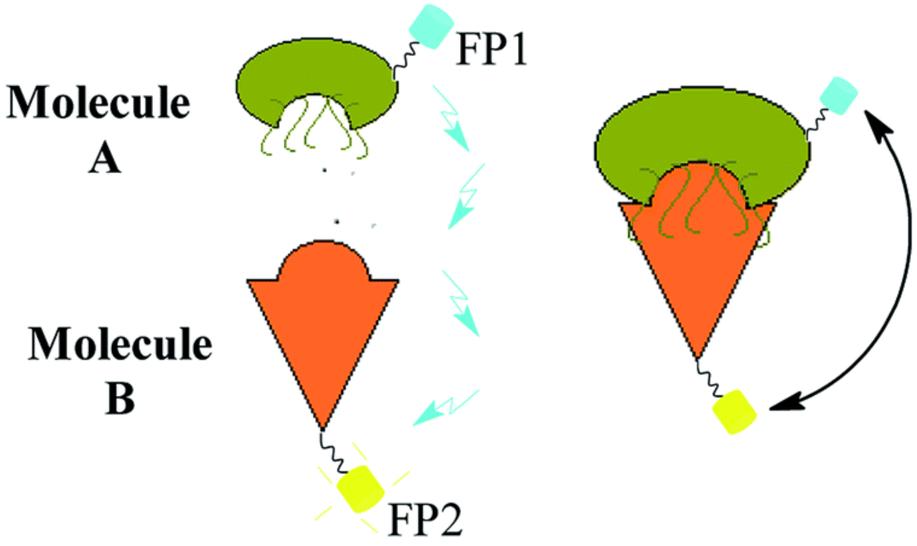


# Scientific Competitions

CASP 14

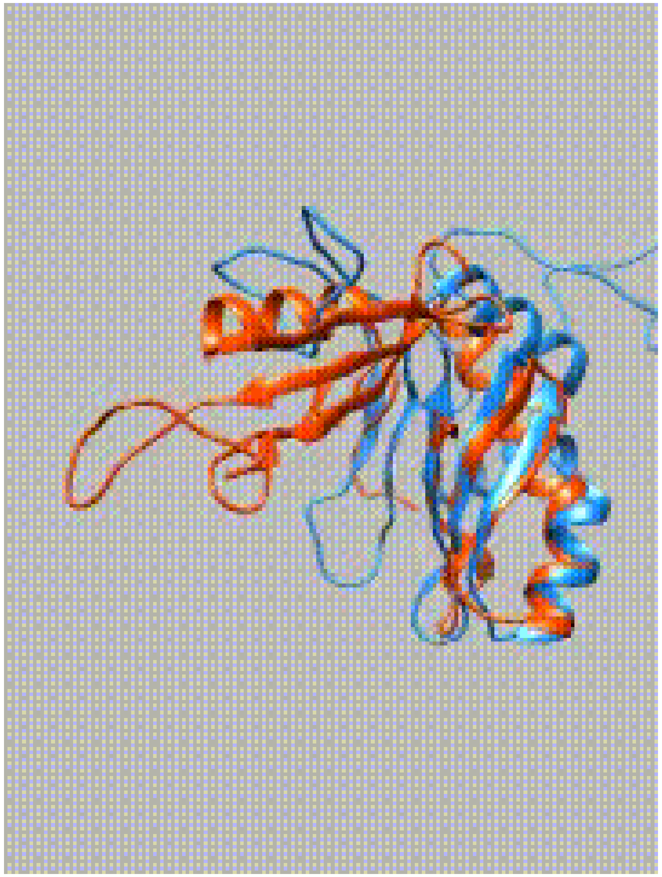
**Data-assisted or hybrid modeling, in which low-resolution experimental data are combined with computational methods, is becoming increasing important for a range of experimental data, including NMR, chemical cross-linking and surface labeling, X-ray and neutron scattering, electron microscopy and FRET.**



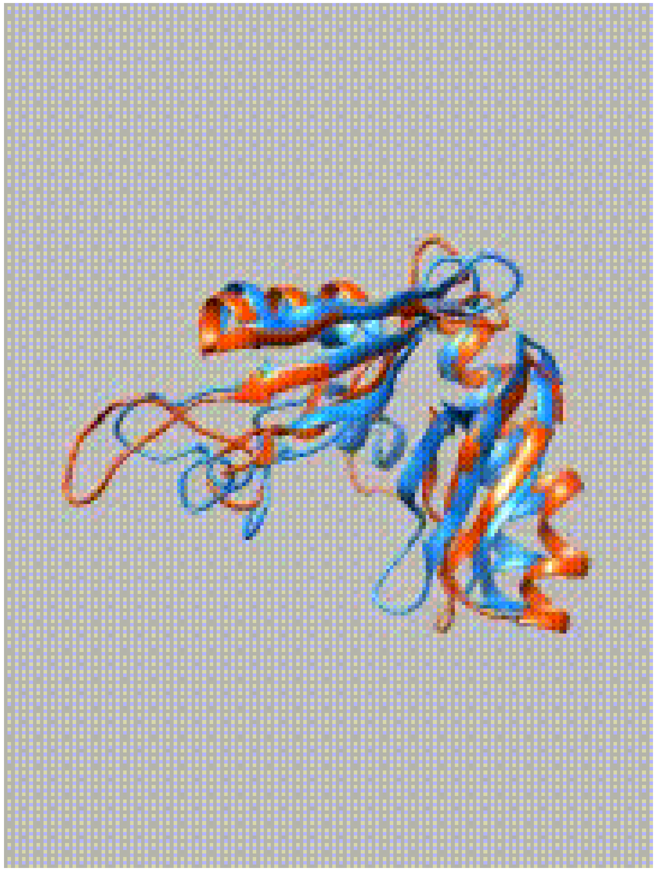


# Scientific Competitions

Data-assisted or hybrid modeling, in which low-resolution experimental data are combined with computational methods, is becoming increasing important for a range of experimental data, including NMR, chemical cross-linking and surface labeling, X-ray and neutron scattering, electron microscopy and FRET

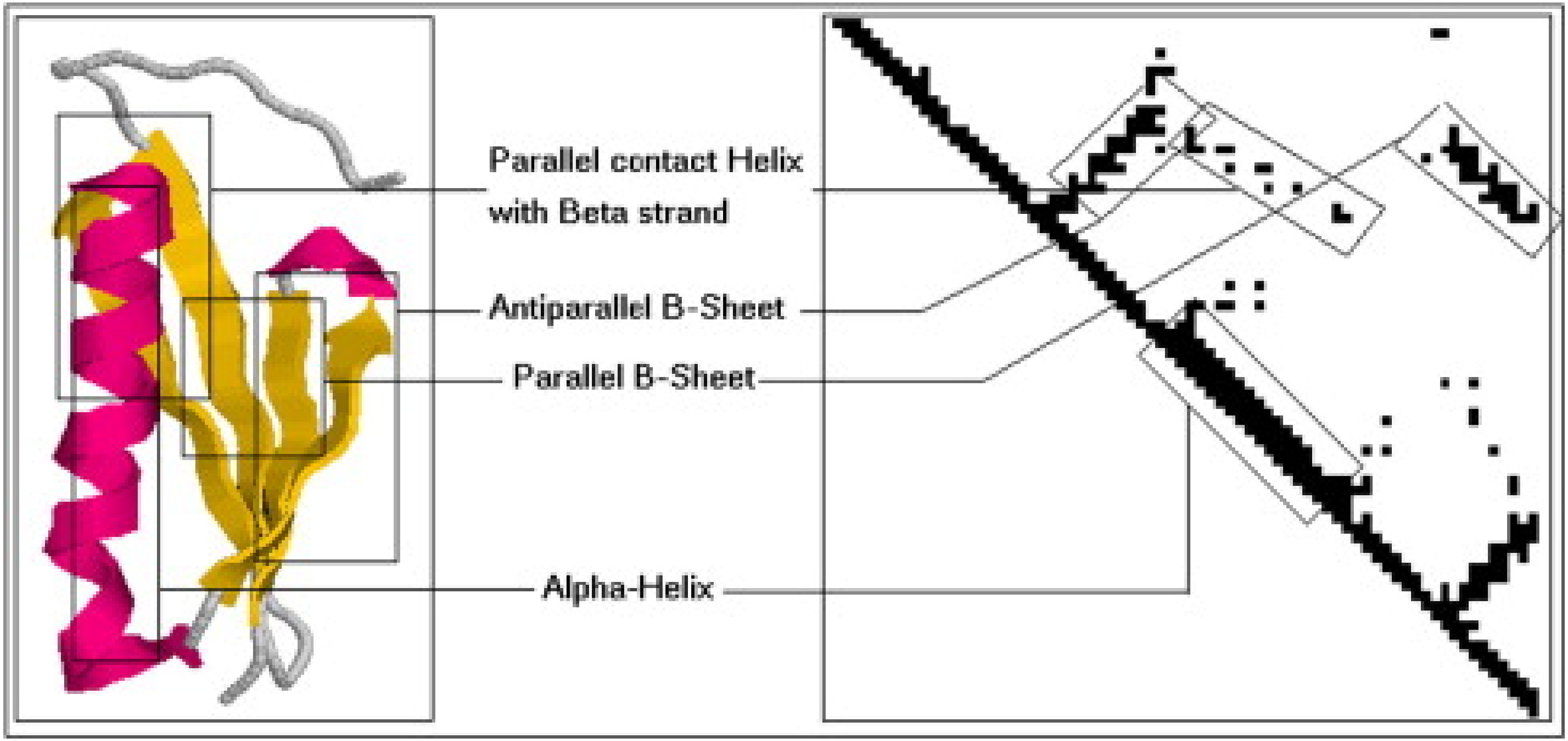


without restrains

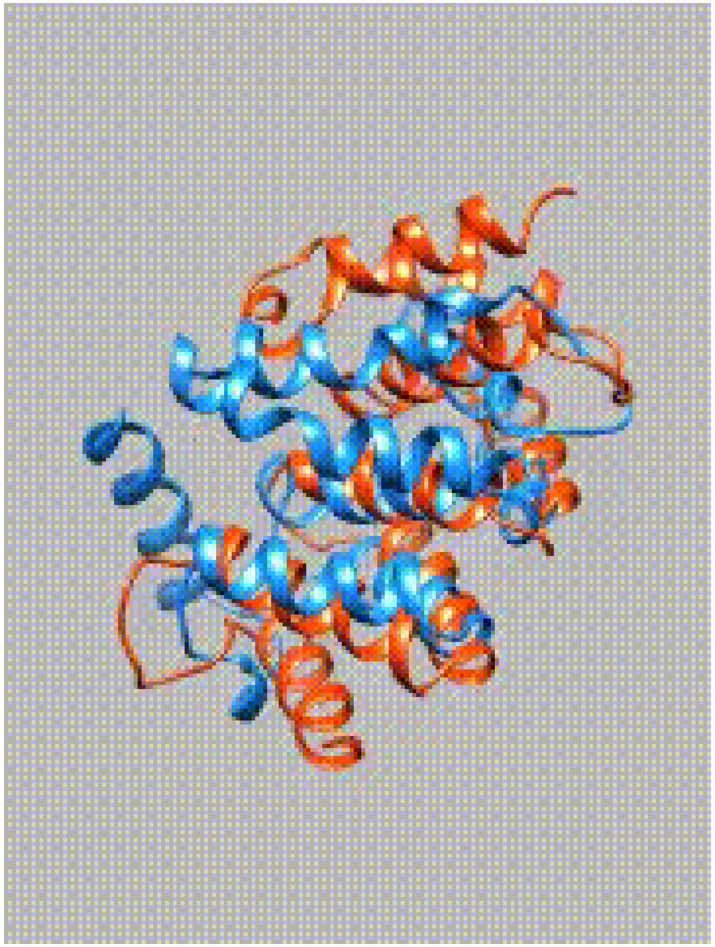


with restrains

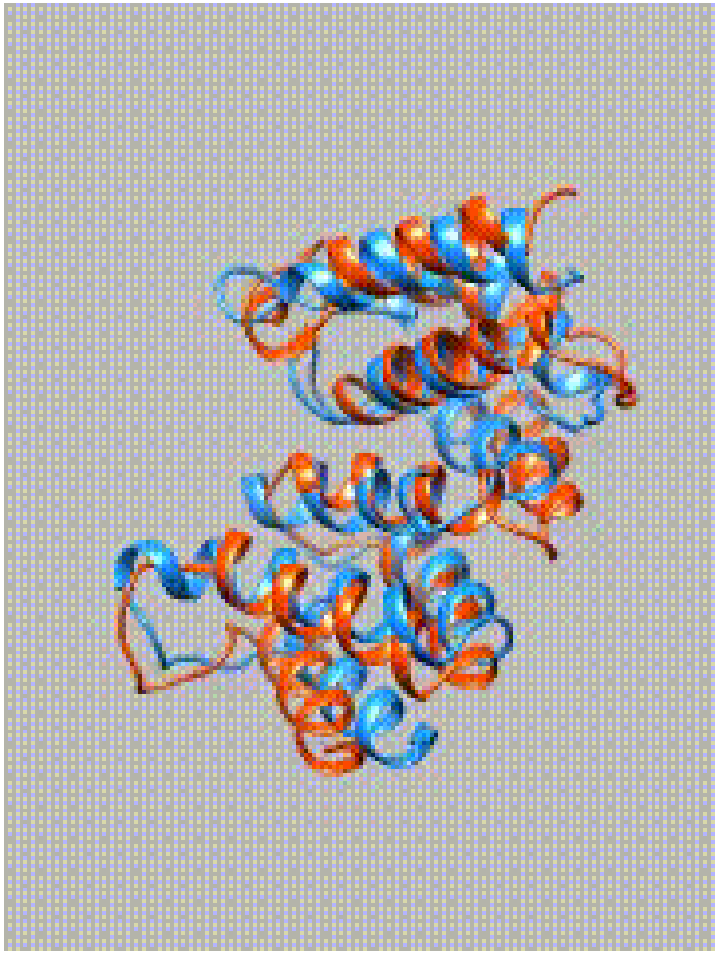
# Residue-residue contact prediction



## Residue-residue contact prediction



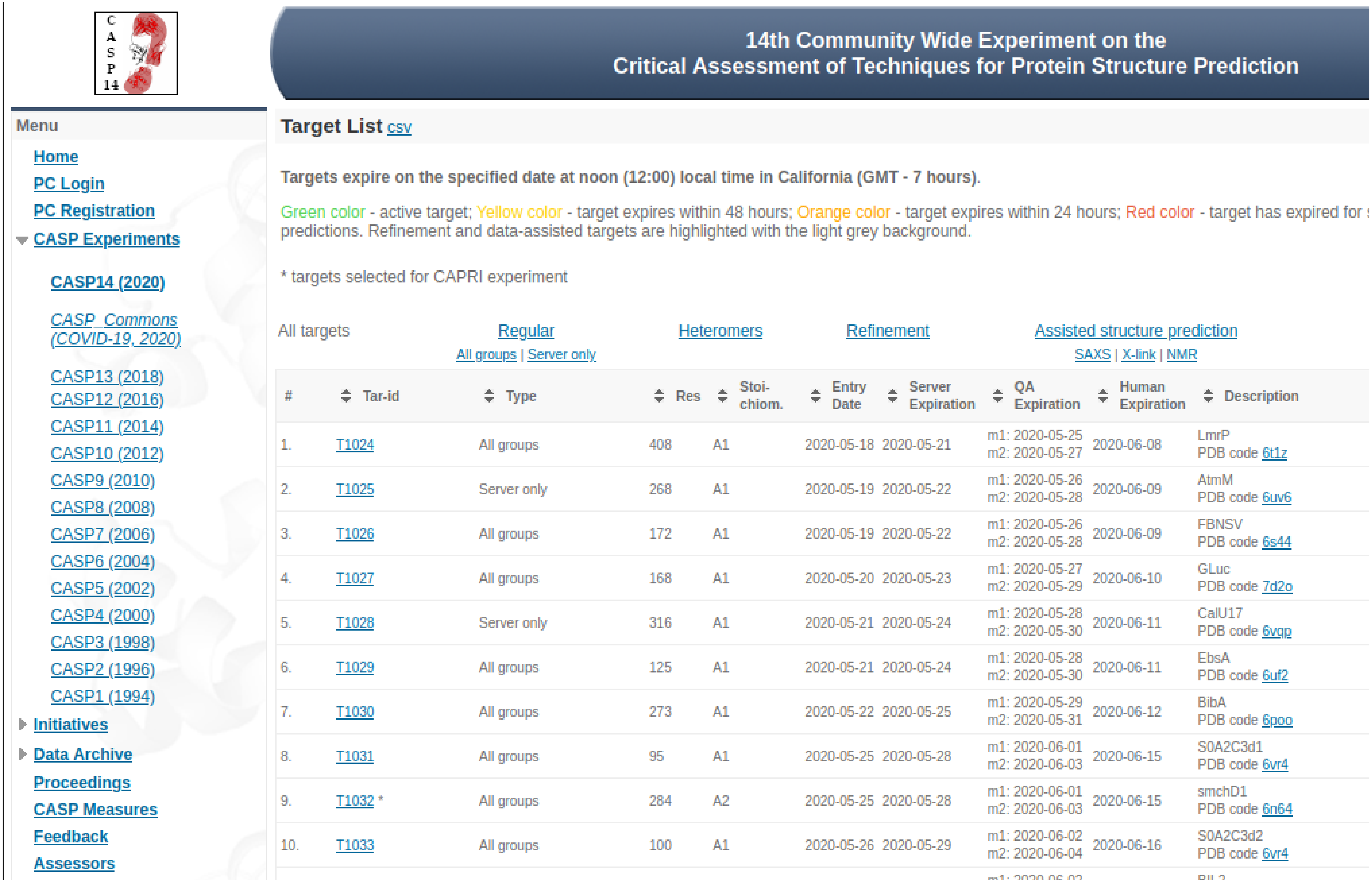
without restrains



with restrains

# Scientific Competitions

ADP



## 14th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction

Menu

- Home
- PC Login
- PC Registration
- CASP Experiments
  - CASP14 (2020)
  - *CASP_Commons (COVID-19, 2020)*
  - CASP13 (2018)
  - CASP12 (2016)
  - CASP11 (2014)
  - CASP10 (2012)
  - CASP9 (2010)
  - CASP8 (2008)
  - CASP7 (2006)
  - CASP6 (2004)
  - CASP5 (2002)
  - CASP4 (2000)
  - CASP3 (1998)
  - CASP2 (1996)
  - CASP1 (1994)
- Initiatives
- Data Archive
- Proceedings
- CASP Measures
- Feedback
- Assessors

### Target List csv

**Targets expire on the specified date at noon (12:00) local time in California (GMT - 7 hours).**

Green color - active target; Yellow color - target expires within 48 hours; Orange color - target expires within 24 hours; Red color - target has expired for predictions. Refinement and data-assisted targets are highlighted with the light grey background.

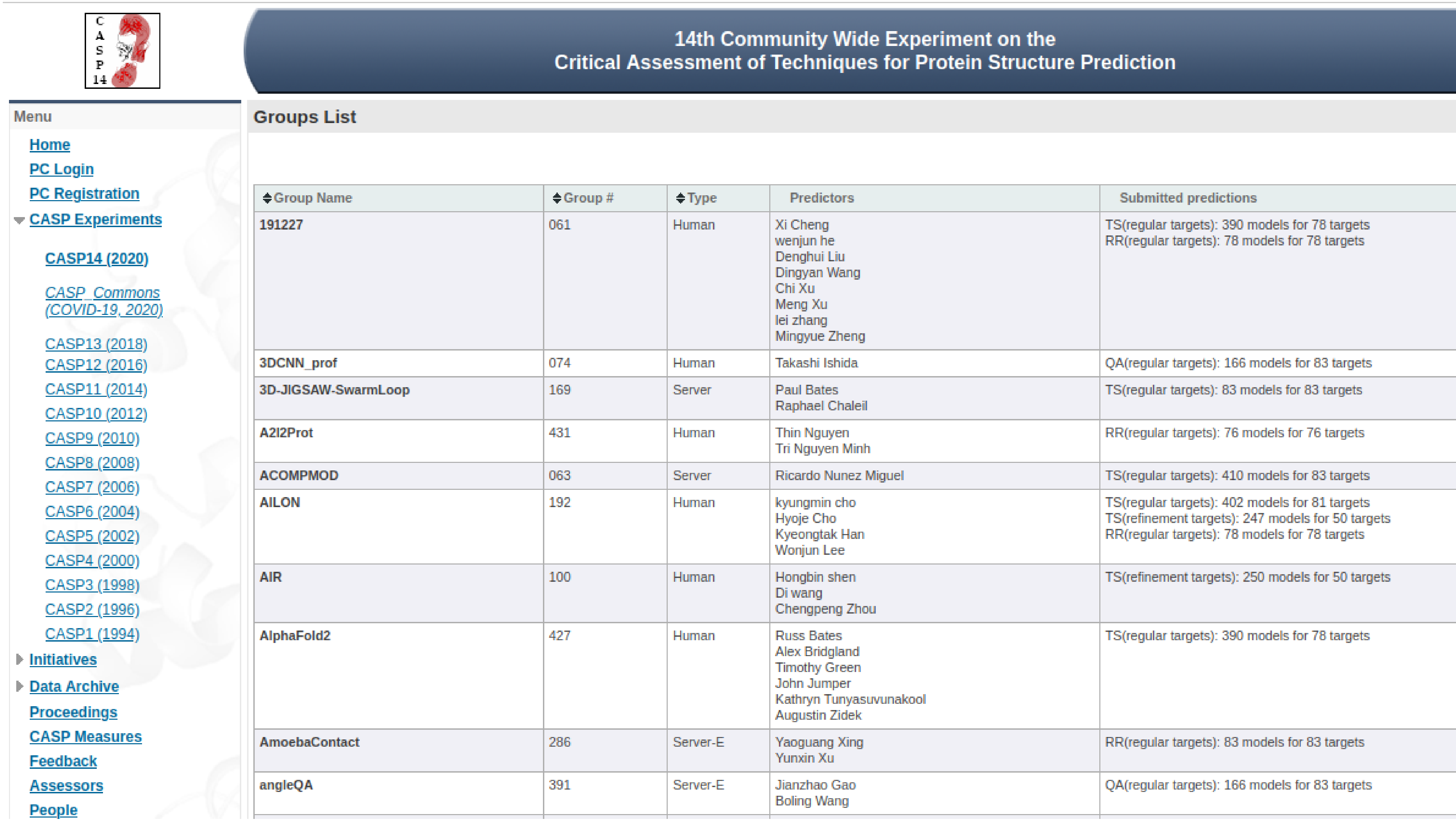
* targets selected for CAPRI experiment

All targets | Regular (All groups | Server only) | Heteromers | Refinement | Assisted structure prediction (SAXS | X-link | NMR)

| # | Tar-id | Type | Res | Stoi-chiom. | Entry Date | Server Expiration | QA Expiration | Human Expiration | Description |
|---|---|---|---|---|---|---|---|---|---|
| 1. | T1024 | All groups | 408 | A1 | 2020-05-18 | 2020-05-21 | m1: 2020-05-25<br>m2: 2020-05-27 | 2020-06-08 | LmrP<br>PDB code 6t1z |
| 2. | T1025 | Server only | 268 | A1 | 2020-05-19 | 2020-05-22 | m1: 2020-05-26<br>m2: 2020-05-28 | 2020-06-09 | AtmM<br>PDB code 6uv6 |
| 3. | T1026 | All groups | 172 | A1 | 2020-05-19 | 2020-05-22 | m1: 2020-05-26<br>m2: 2020-05-28 | 2020-06-09 | FBNSV<br>PDB code 6s44 |
| 4. | T1027 | All groups | 168 | A1 | 2020-05-20 | 2020-05-23 | m1: 2020-05-27<br>m2: 2020-05-29 | 2020-06-10 | GLuc<br>PDB code 7d2o |
| 5. | T1028 | Server only | 316 | A1 | 2020-05-21 | 2020-05-24 | m1: 2020-05-28<br>m2: 2020-05-30 | 2020-06-11 | CalU17<br>PDB code 6vqp |
| 6. | T1029 | All groups | 125 | A1 | 2020-05-21 | 2020-05-24 | m1: 2020-05-28<br>m2: 2020-05-30 | 2020-06-11 | EbsA<br>PDB code 6uf2 |
| 7. | T1030 | All groups | 273 | A1 | 2020-05-22 | 2020-05-25 | m1: 2020-05-29<br>m2: 2020-05-31 | 2020-06-12 | BibA<br>PDB code 6poo |
| 8. | T1031 | All groups | 95 | A1 | 2020-05-25 | 2020-05-28 | m1: 2020-06-01<br>m2: 2020-06-03 | 2020-06-15 | S0A2C3d1<br>PDB code 6vr4 |
| 9. | T1032 * | All groups | 284 | A2 | 2020-05-25 | 2020-05-28 | m1: 2020-06-01<br>m2: 2020-06-03 | 2020-06-15 | smchD1<br>PDB code 6n64 |
| 10. | T1033 | All groups | 100 | A1 | 2020-05-26 | 2020-05-29 | m1: 2020-06-02<br>m2: 2020-06-04 | 2020-06-16 | S0A2C3d2<br>PDB code 6vr4 |

# Scientific Competitions

ADP



14th Community Wide Experiment on the
Critical Assessment of Techniques for Protein Structure Prediction

## Groups List

| Group Name | Group # | Type | Predictors | Submitted predictions |
|---|---|---|---|---|
| **191227** | 061 | Human | Xi Cheng<br>wenjun he<br>Denghui Liu<br>Dingyan Wang<br>Chi Xu<br>Meng Xu<br>lei zhang<br>Mingyue Zheng | TS(regular targets): 390 models for 78 targets<br>RR(regular targets): 78 models for 78 targets |
| **3DCNN_prof** | 074 | Human | Takashi Ishida | QA(regular targets): 166 models for 83 targets |
| **3D-JIGSAW-SwarmLoop** | 169 | Server | Paul Bates<br>Raphael Chaleil | TS(regular targets): 83 models for 83 targets |
| **A2I2Prot** | 431 | Human | Thin Nguyen<br>Tri Nguyen Minh | RR(regular targets): 76 models for 76 targets |
| **ACOMPMOD** | 063 | Server | Ricardo Nunez Miguel | TS(regular targets): 410 models for 83 targets |
| **AILON** | 192 | Human | kyungmin cho<br>Hyoje Cho<br>Kyeongtak Han<br>Wonjun Lee | TS(regular targets): 402 models for 81 targets<br>TS(refinement targets): 247 models for 50 targets<br>RR(regular targets): 78 models for 78 targets |
| **AIR** | 100 | Human | Hongbin shen<br>Di wang<br>Chengpeng Zhou | TS(refinement targets): 250 models for 50 targets |
| **AlphaFold2** | 427 | Human | Russ Bates<br>Alex Bridgland<br>Timothy Green<br>John Jumper<br>Kathryn Tunyasuvunakool<br>Augustin Zidek | TS(regular targets): 390 models for 78 targets |
| **AmoebaContact** | 286 | Server-E | Yaoguang Xing<br>Yunxin Xu | RR(regular targets): 83 models for 83 targets |
| **angleQA** | 391 | Server-E | Jianzhao Gao<br>Boling Wang | QA(regular targets): 166 models for 83 targets |

## Critical Assessment of Techniques for Protein Structure Prediction (CASP)

**Evaluation of the results is carried out in the following prediction catego**

- tertiary structure prediction (all CASPs)
- secondary structure prediction (dropped after CASP5)
- prediction of structure complexes (CASP2 only;
  a separate experiment - CAPRI—carries on this subject)
- residue-residue contact prediction (starting CASP4)
- disordered regions prediction (starting CASP5)
- domain boundary prediction (CASP6–CASP8)
- function prediction (starting CASP6)
- model quality assessment (starting CASP7)
- model refinement (starting CASP7)
- high-accuracy template-based prediction (starting CASP7)

# Critical Assessment of Techniques for Protein Structure Prediction (CASP)

**Evaluation of the results is carried out in the following prediction catego**

C
A
S
P
14

- **tertiary structure prediction (all CASPs)**
- secondary structure prediction (dropped after CASP5)
- prediction of structure complexes (CASP2 only;
 a separate experiment - CAPRI—carries on this subject)
- re due contact prediction (st ASP4)
- d egions prediction (startin )
- d ndary prediction (CASP6-
- fu diction (starting CASP6)
- model quality assessment (starting CASP7)
- m ment (starting CASP7)
- late-based pr g CASP7)

Janusz Bujnicki

Andrzej Koliński

International Institute of Molecular and Cell Biology in Warsaw

WCh
Wydział Chemii
Uniwersytet Warszawski

CASP 14

## Critical Assessment of Techniques for Protein Structure Prediction (CASP)

**Evaluation of the results is carried out in the following prediction catego**

- tertiary structure prediction (all CASPs)
- secondary structure prediction (dropped after CASP5)
- prediction of structure complexes (CASP2 only;
  a separate experiment - CAPRI—carries on this su
- **residue-residue contact prediction**
- disordered regions prediction (starting CASP5)
- domain boundary prediction (CASP6–CASP8)
- function prediction (starting CASP6)
- model quality assessment (starting CASP7)
- model refinement (starting CASP7)
- high-accuracy template-based prediction

Michał Piętal

International Institute of Molecular and Cell Biology in Warsaw

POLITECHNIKA RZESZOWSKA im. IGNACEGO ŁUKASIEWICZA

# Critical Assessment of Techniques for Protein Structure Prediction (CASP)

CASP 14

**Evaluation of the results is carried out in the following prediction catego**

- tertiary structure prediction (all CASPs)
- secondary structure prediction (dropped after
- prediction of structure complexes (CASP2 onl

  a separate experiment - CAPRI—carries on this
- residue-residue contact prediction
- **disordered regions prediction** (starting CASP5
- domain boundary prediction (CASP6–CASP8)
- function prediction (starting CASP6)
- model quality assessment (starting CASP
- model refinement (starting CASP7)
- high-accuracy template-based prediction (

Łukasz P. Kozłowski

International Institute of Molecular and Cell Biology in Warsaw

WYDZIAŁ MATEMATYKI, INFORMATYKI I MECHANIKI

MIM

CASP 14

# Critical Assessment of Techniques for Protein Structure Prediction (CASP)

**Evaluation of the results is carried out in the following prediction catego**

- tertiary structure prediction (all CASPs)
- secondary structure prediction (dropped a
- prediction of structure complexes (CASP2

  a separate experiment - CAPRI—carries on t
- residue-residue contact prediction
- disordered regions prediction (starting CASP5)
- domain boundary prediction (CASP6–CASP8)
- function prediction (starting CASP6)
- **model quality assessment**
- model refinement (starting CASP7)
- high-accuracy template-based predic

Marcin Pawłowski

INTERNATIONAL INSTITUTE OF MOLECULAR AND CELL BIOLOGY • IN WARSAW •

International Institute of Molecular and Cell Biology in Warsaw

precisionlife
driving precision medicine

# Scientific Competitions

DEEP LEARNING

TensorFlow

Keras

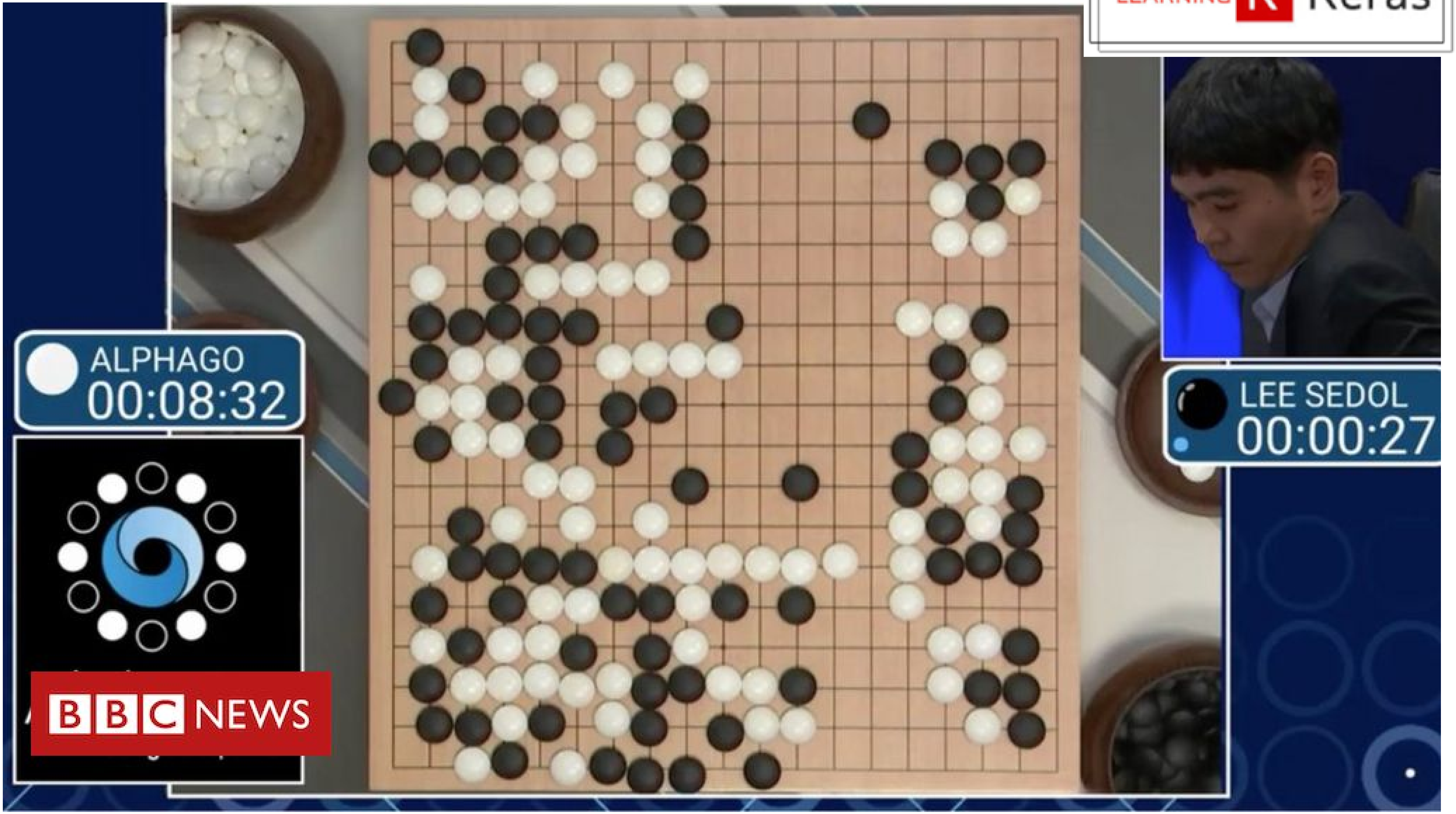
# Scientific Competitions

AlphaGo

DEEP LEARNING

TensorFlow

Keras



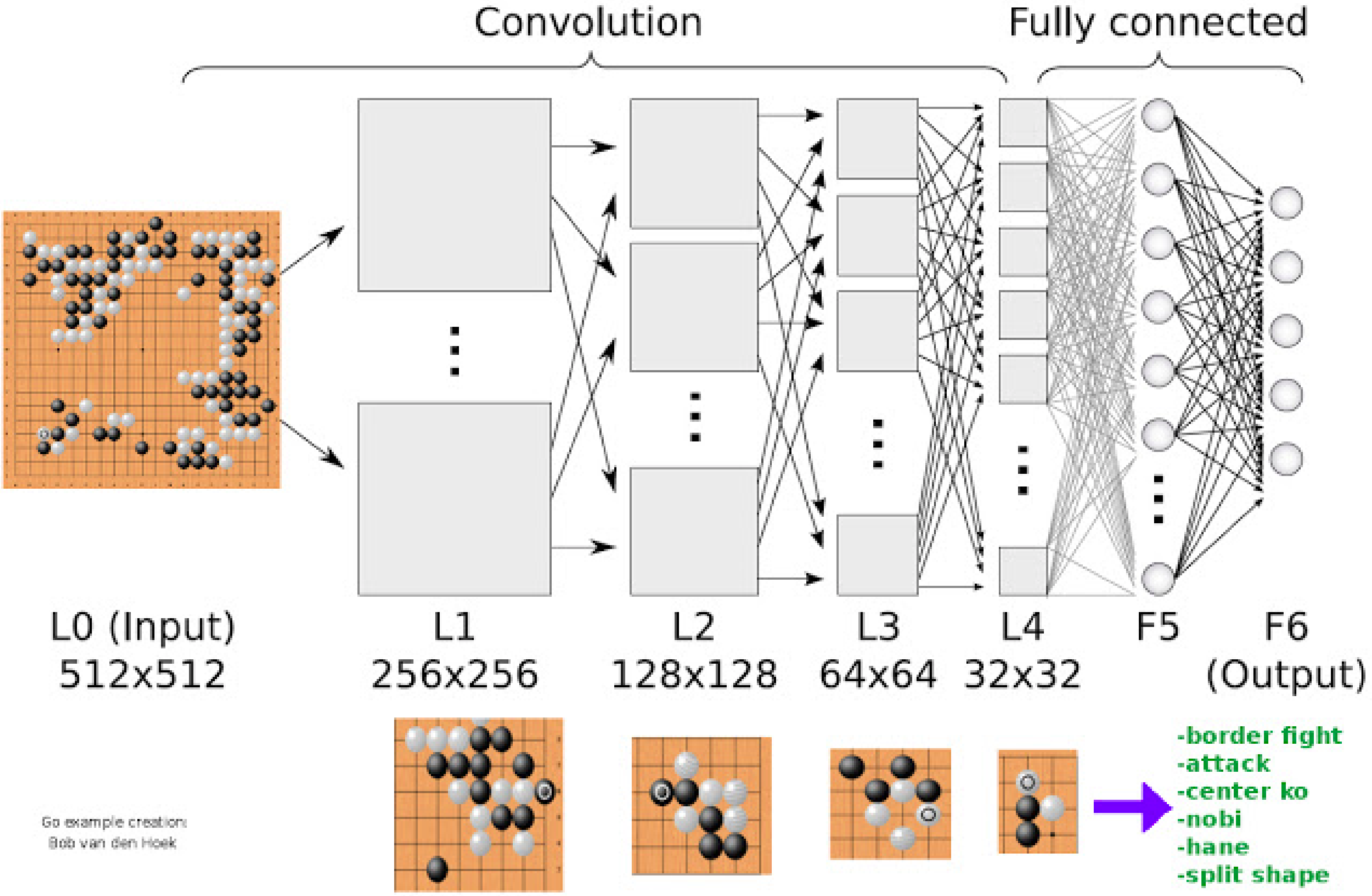
# Scientific Competitions

AlphaGo

DEEP LEARNING

TensorFlow

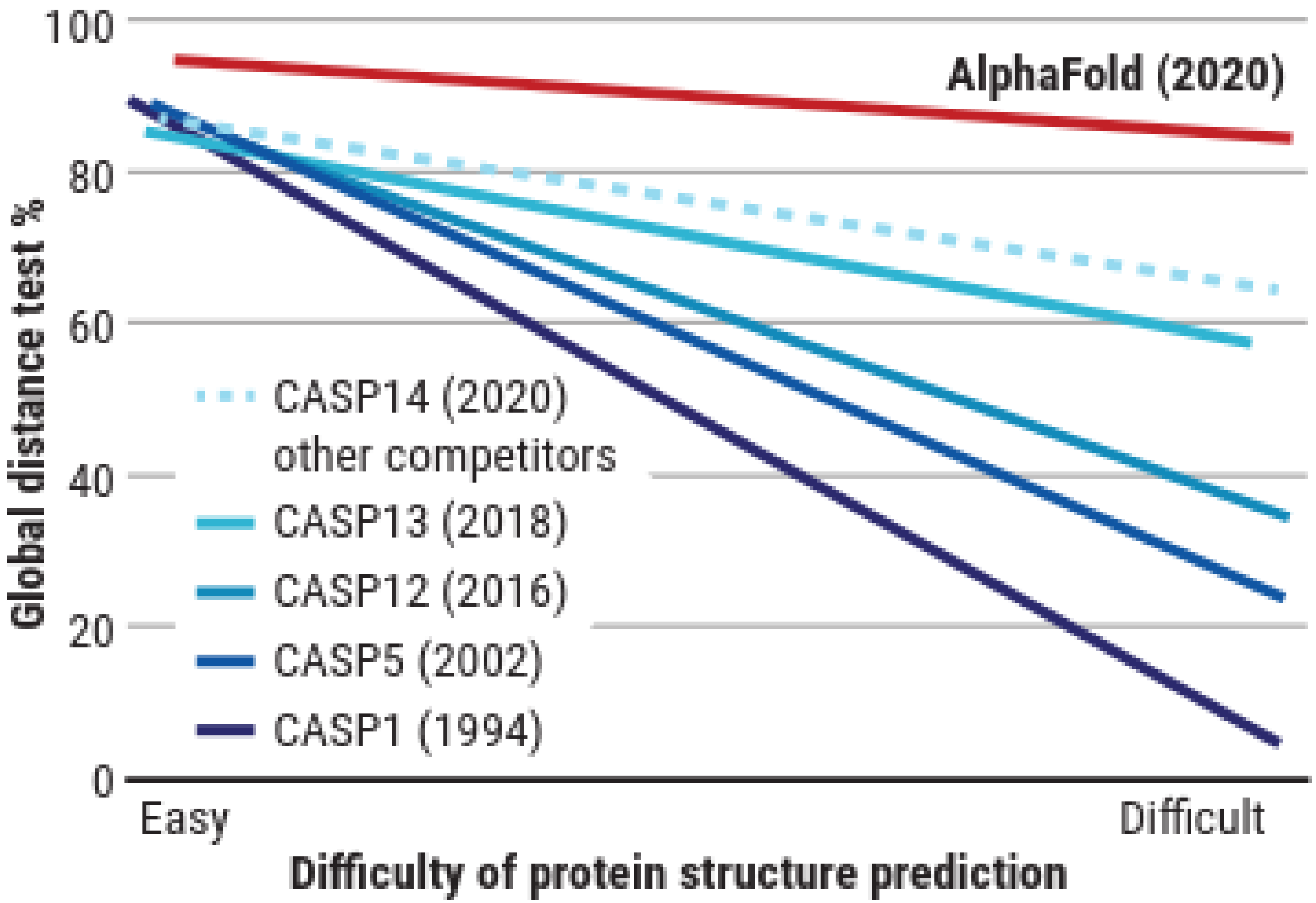
Keras



-border fight
-attack
-center ko
-nobi
-hane
-split shape

Go example creation:
Bob van den Hoek

# AlphaFold

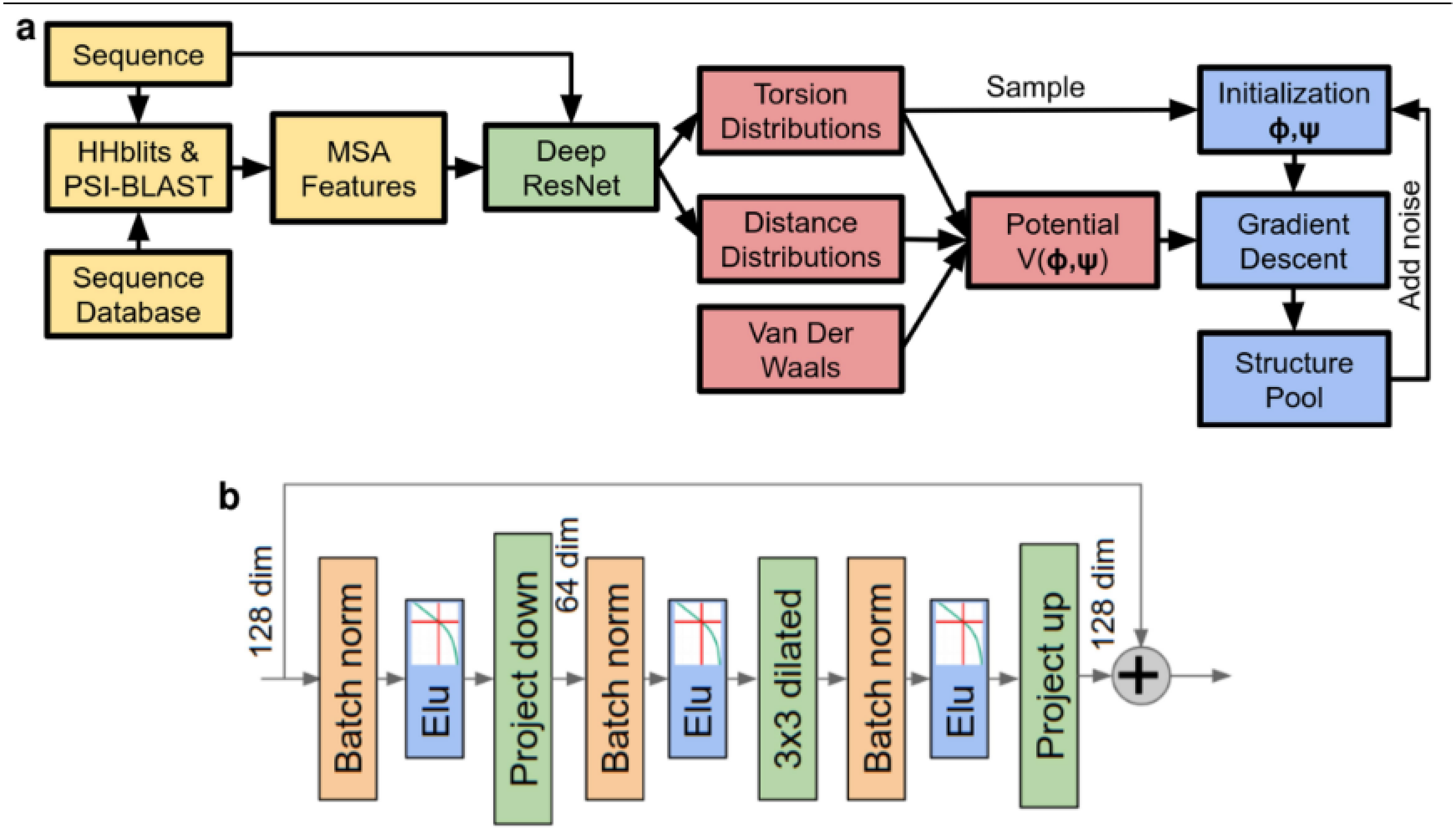

100
80
60
40
20
0
Global distance test %
AlphaFold (2020)
CASP14 (2020) other competitors
CASP13 (2018)
CASP12 (2016)
CASP5 (2002)
CASP1 (1994)
Easy
Difficult
Difficulty of protein structure prediction

# Presentations & Videos from CASP15

https://predictioncenter.org/casp15/doc/presentations/

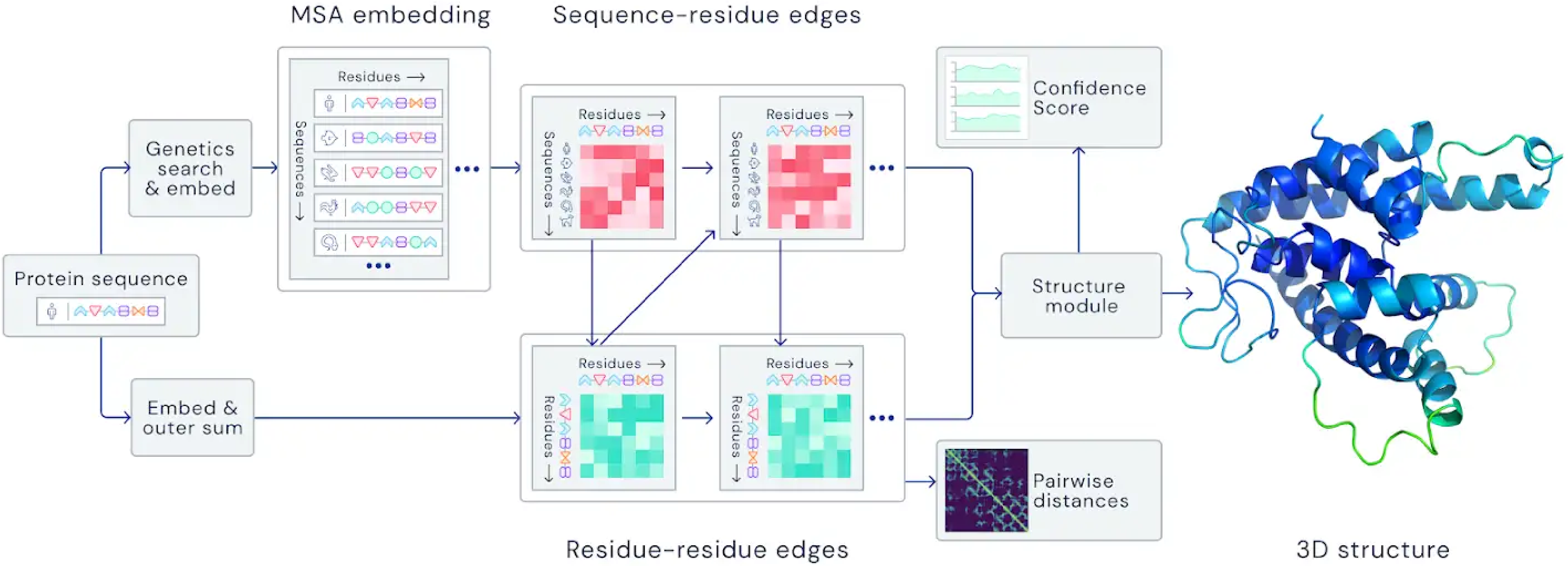
https://www.youtube.com/@CASP-Prediction-Center/videos

# AlphaFold

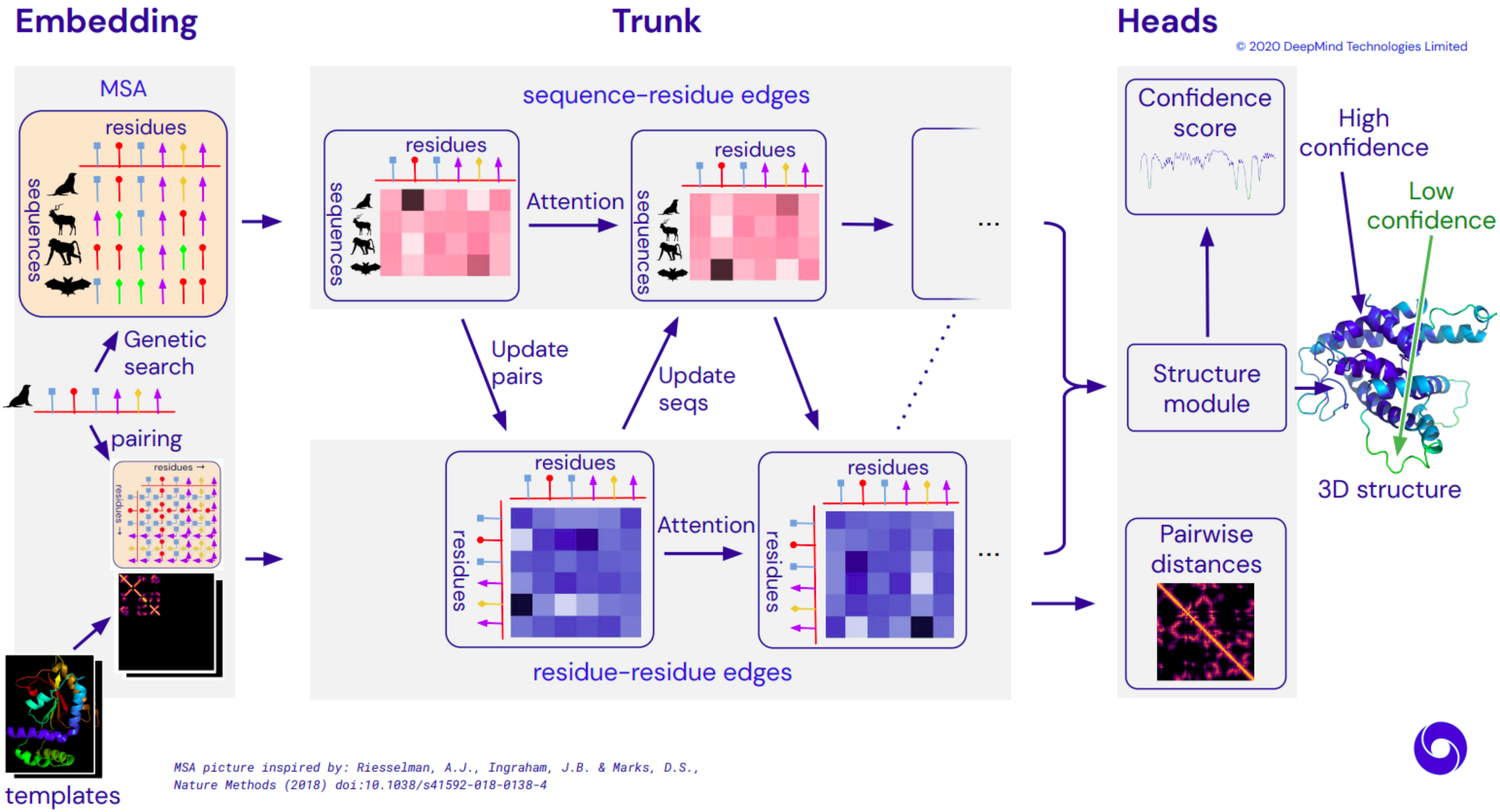


**Extended Data Fig. 1 | Schematics of the folding system and neural network.** **a**, The overall folding system. Feature extraction stages (constructing the MSA using sequence database search and computing MSA-based features) are shown in yellow; the structure-prediction neural network in green; potential construction in red; and structure realization in blue. **b**, The layers used in one block of the deep residual convolutional network. The dilated convolution is applied to activations of reduced dimension. The output of the block is added to the representation from the previous layer. The bypass connections of the residual network enable gradients to pass back through the network undiminished, permitting the training of very deep networks.

# AlphaFold


MSA embedding
Sequence-residue edges
Genetics search & embed
Residues →
Sequences →
Protein sequence
Embed & outer sum
Confidence Score
Structure module
Pairwise distances
Residue-residue edges
3D structure

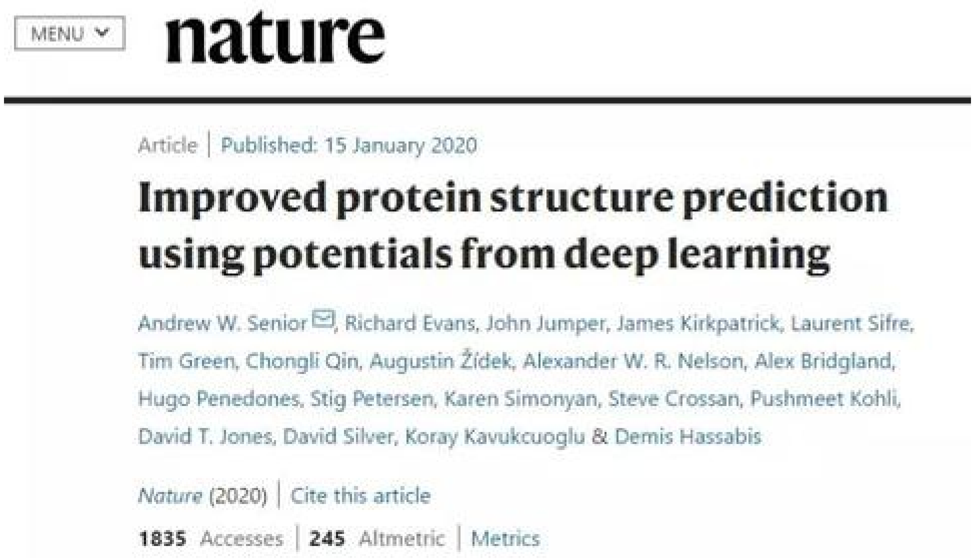
# AlphaFold2



MSA picture inspired by: Riesselman, A.J., Ingraham, J.B. & Marks, D.S., Nature Methods (2018) doi:10.1038/s41592-018-0138-4

https://predictioncenter.org/casp14/doc/presentations/2020_12_01_TS_predictor_AlphaFold2.pdf

## AlphaFold



Article | Published: 15 January 2020

### Improved protein structure prediction using potentials from deep learning

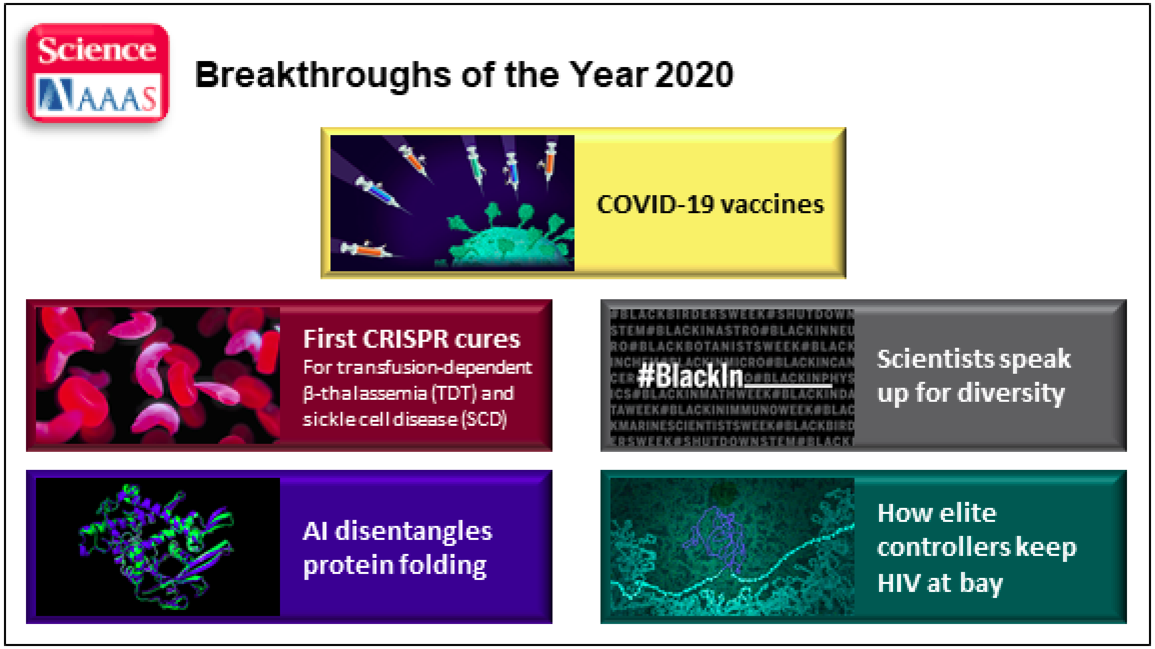
Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu & Demis Hassabis

*Nature* (2020) | Cite this article

**1835** Accesses | **245** Altmetric | Metrics

# Scientific Competitions

## AlphaFold



Breakthroughs of the Year 2020

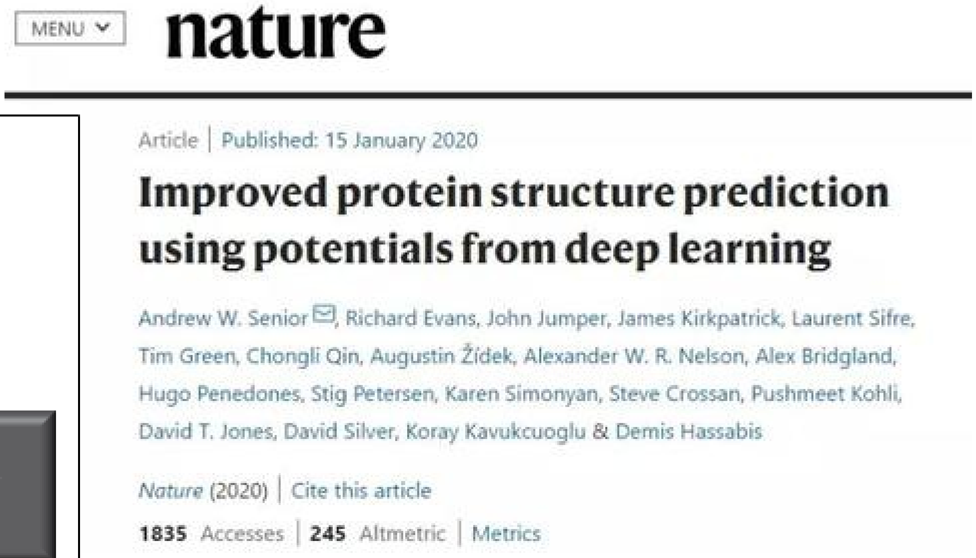
COVID-19 vaccines

First CRISPR cures
For transfusion-dependent β-thalassemia (TDT) and sickle cell disease (SCD)

Scientists speak up for diversity

AI disentangles protein folding

How elite controllers keep HIV at bay



Article | Published: 15 January 2020

Improved protein structure prediction using potentials from deep learning

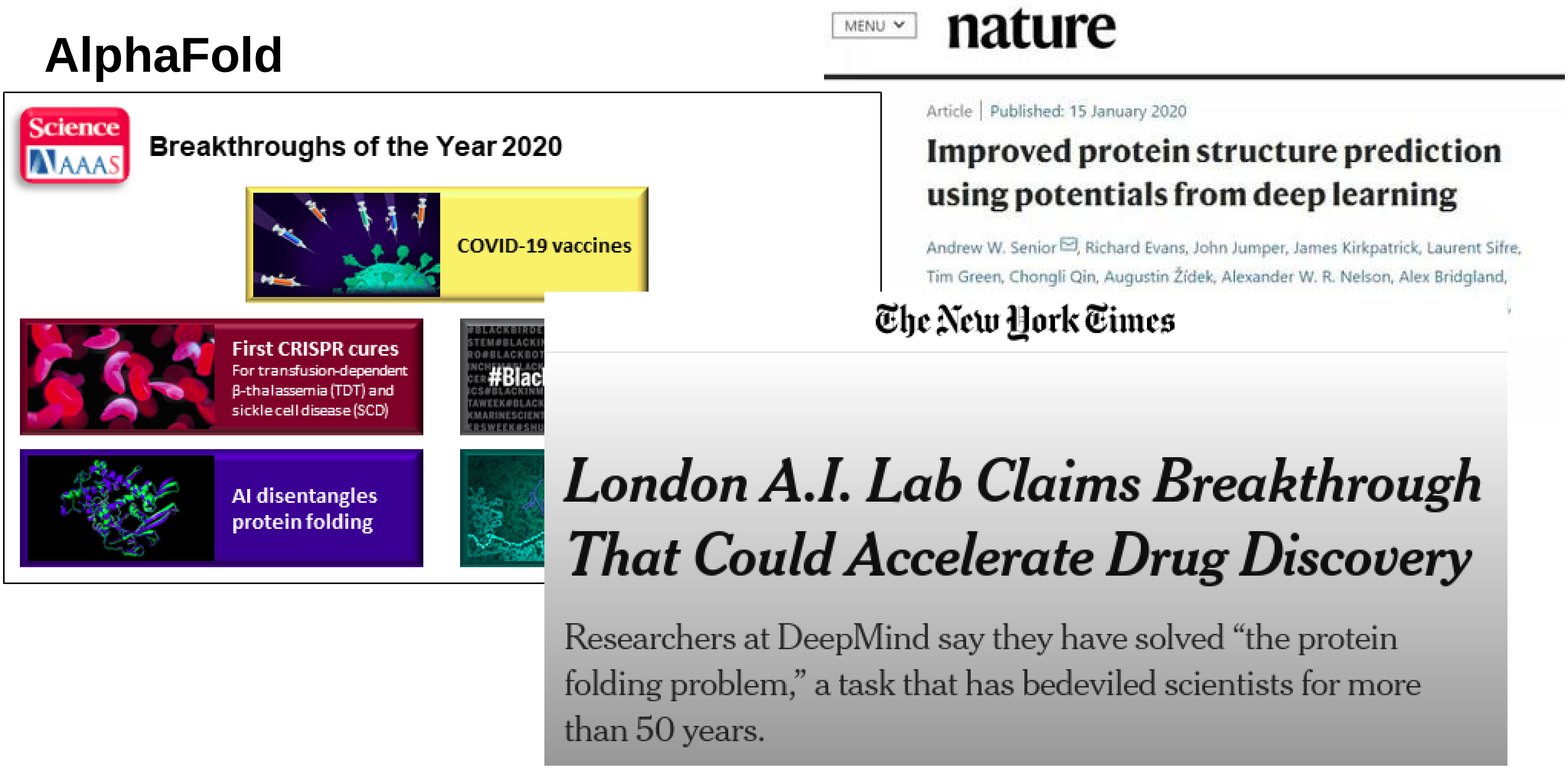
Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu & Demis Hassabis

*Nature* (2020) | Cite this article

1835 Accesses | 245 Altmetric | Metrics

# Scientific Competitions

## AlphaFold



Breakthroughs of the Year 2020

COVID-19 vaccines

First CRISPR cures
For transfusion-dependent β-thalassemia (TDT) and sickle cell disease (SCD)

AI disentangles protein folding

MENU

nature

Article | Published: 15 January 2020

Improved protein structure prediction using potentials from deep learning

Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland,

The New York Times

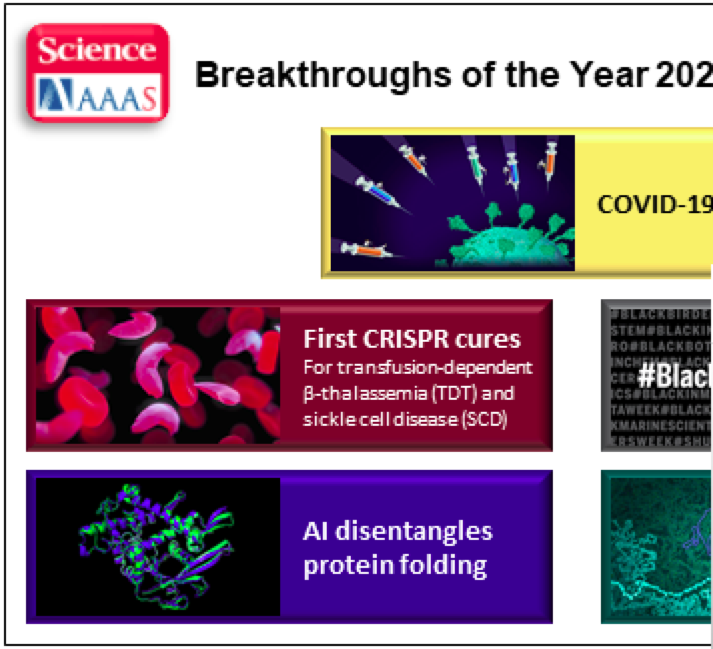
*London A.I. Lab Claims Breakthrough That Could Accelerate Drug Discovery*

Researchers at DeepMind say they have solved “the protein folding problem,” a task that has bedeviled scientists for more than 50 years.

# Scientific Competitions

ADP

## AlphaFold



Science AAAS

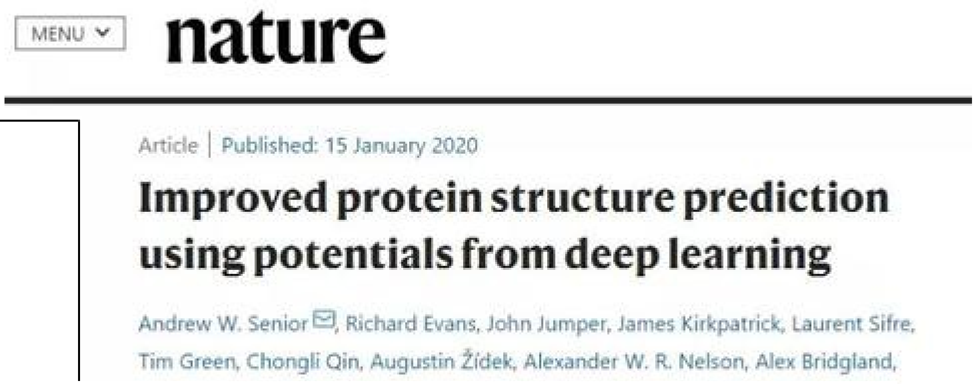
Breakthroughs of the Year 2020

COVID-19 vaccines

First CRISPR cures
For transfusion-dependent β-thalassemia (TDT) and sickle cell disease (SCD)

#Blac

AI disentangles protein folding



nature

Article | Published: 15 January 2020

Improved protein structure prediction using potentials from deep learning

Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland,

The New York Times

*London A.I. Lab Claims Breakthrough That Could Accelerate Drug Discovery*

Researchers at DeepMind say they have solved "the protein folding problem," a task that has bedeviled scientists for more than 50 years.

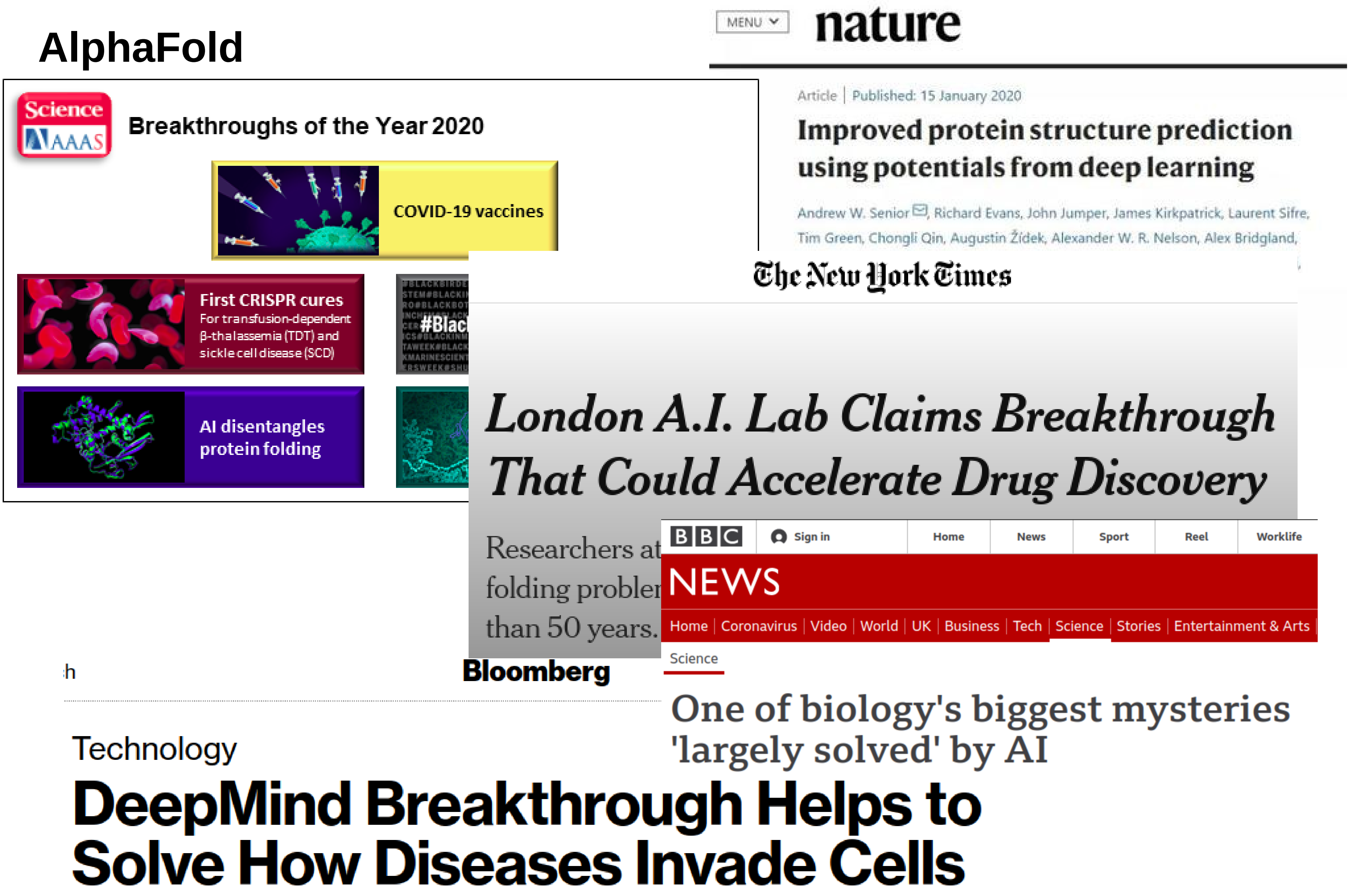
h Bloomberg

Technology

**DeepMind Breakthrough Helps to Solve How Diseases Invade Cells**

# Scientific Competitions

## AlphaFold



Science AAAS

Breakthroughs of the Year 2020

COVID-19 vaccines

First CRISPR cures
For transfusion-dependent β-thalassemia (TDT) and sickle cell disease (SCD)

#Blac

AI disentangles protein folding

MENU nature

Article | Published: 15 January 2020

**Improved protein structure prediction using potentials from deep learning**

Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland,

The New York Times

*London A.I. Lab Claims Breakthrough That Could Accelerate Drug Discovery*

Researchers at
folding probler
than 50 years.

BBC | Sign in | Home | News | Sport | Reel | Worklife

NEWS

Home | Coronavirus | Video | World | UK | Business | Tech | Science | Stories | Entertainment & Arts |

Science

**One of biology's biggest mysteries 'largely solved' by AI**

h

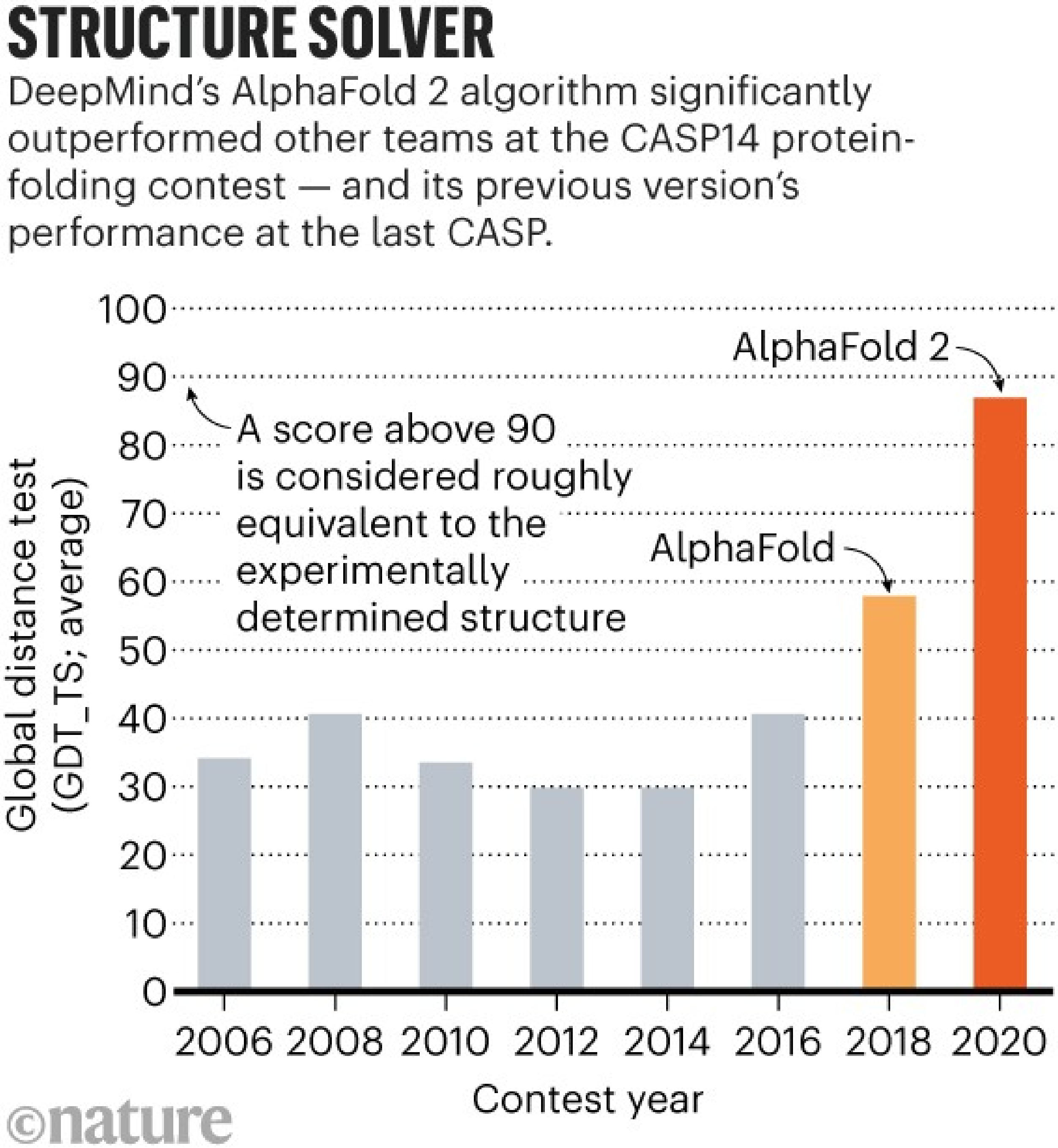
Bloomberg

Technology

**DeepMind Breakthrough Helps to Solve How Diseases Invade Cells**

# Scientific Competitions

## AlphaFold

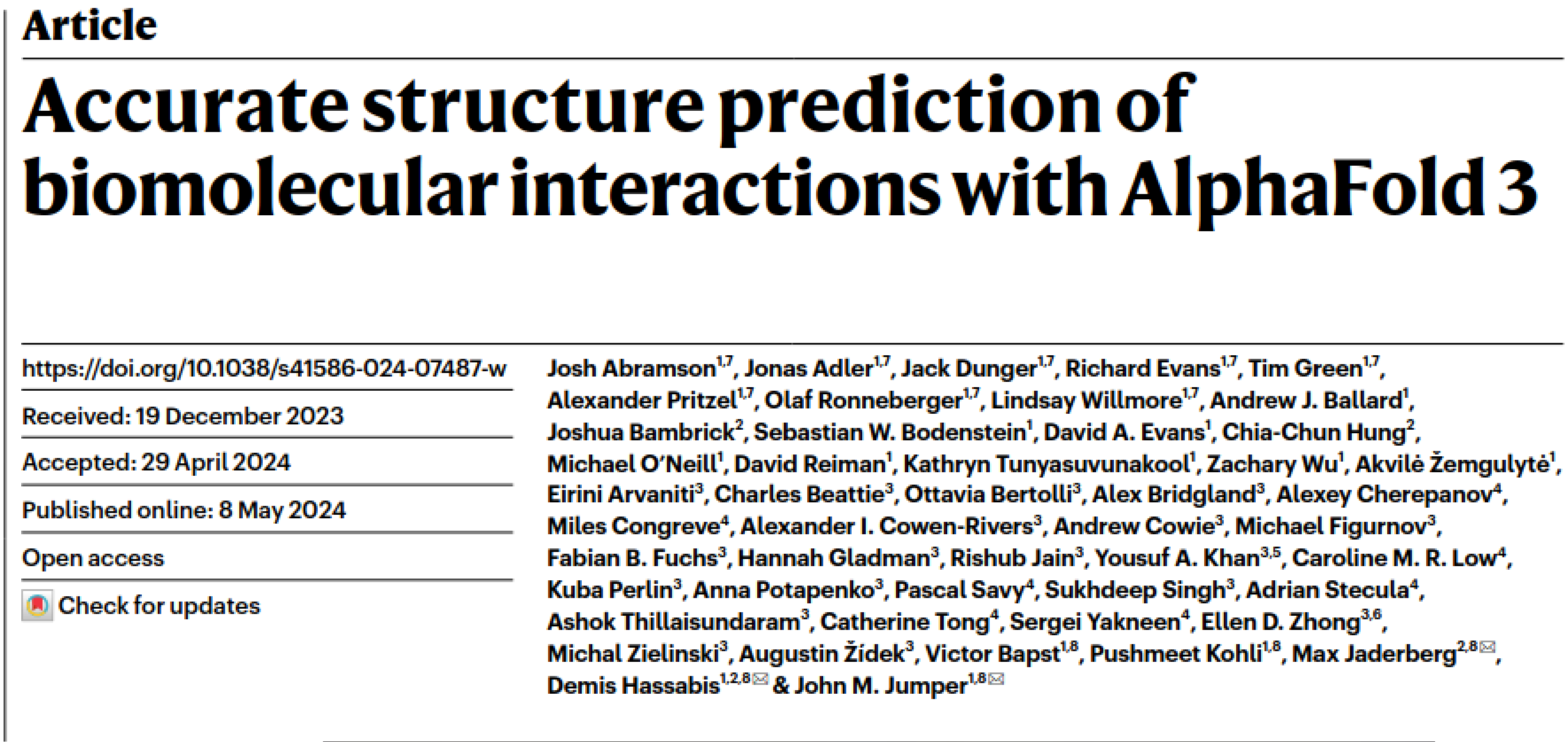


STRUCTURE SOLVER

DeepMind's AlphaFold 2 algorithm significantly outperformed other teams at the CASP14 protein-folding contest — and its previous version's performance at the last CASP.

A score above 90 is considered roughly equivalent to the experimentally determined structure

©nature



Article

# Accurate structure prediction of biomolecular interactions with AlphaFold 3

https://doi.org/10.1038/s41586-024-07487-w

Received: 19 December 2023

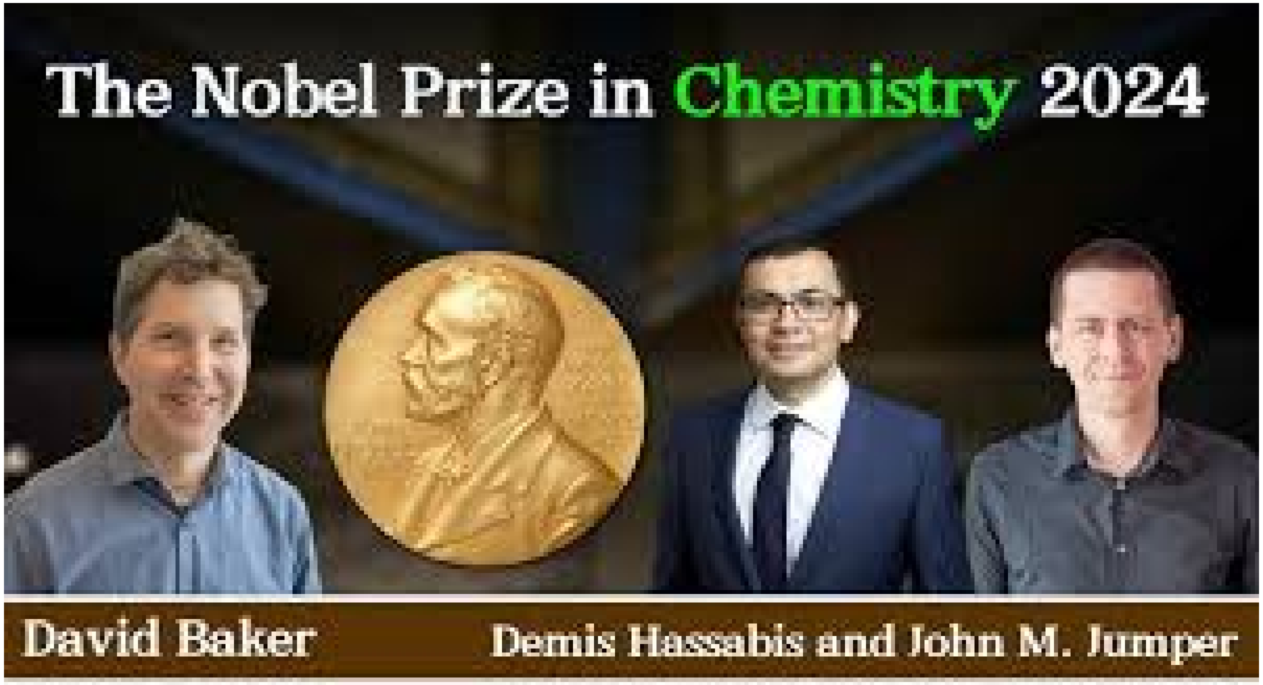
Accepted: 29 April 2024

Published online: 8 May 2024

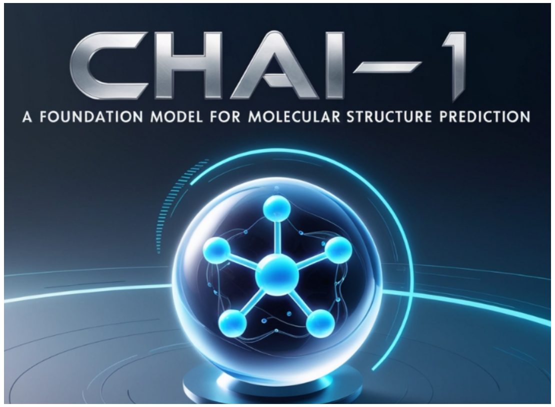
Open access

Check for updates

Josh Abramson[1,7], Jonas Adler[1,7], Jack Dunger[1,7], Richard Evans[1,7], Tim Green[1,7], Alexander Pritzel[1,7], Olaf Ronneberger[1,7], Lindsay Willmore[1,7], Andrew J. Ballard[1], Joshua Bambrick[2], Sebastian W. Bodenstein[1], David A. Evans[1], Chia-Chun Hung[2], Michael O'Neill[1], David Reiman[1], Kathryn Tunyasuvunakool[1], Zachary Wu[1], Akvilė Žemgulytė[1], Eirini Arvaniti[3], Charles Beattie[3], Ottavia Bertolli[3], Alex Bridgland[3], Alexey Cherepanov[4], Miles Congreve[4], Alexander I. Cowen-Rivers[3], Andrew Cowie[3], Michael Figurnov[3], Fabian B. Fuchs[3], Hannah Gladman[3], Rishub Jain[3], Yousuf A. Khan[3,5], Caroline M. R. Low[4], Kuba Perlin[3], Anna Potapenko[3], Pascal Savy[4], Sukhdeep Singh[3], Adrian Stecula[4], Ashok Thillaisundaram[3], Catherine Tong[4], Sergei Yakneen[4], Ellen D. Zhong[3,6], Michal Zielinski[3], Augustin Žídek[3], Victor Bapst[1,8], Pushmeet Kohli[1,8], Max Jaderberg[2,8]✉, Demis Hassabis[1,2,8]✉ & John M. Jumper[1,8]✉


The Nobel Prize in Chemistry 2024
David Baker
Demis Hassabis and John M. Jumper

HelixFold3



CHAI-1

A FOUNDATION MODEL FOR MOLECULAR STRUCTURE PREDICTION

ANKH

ESM3

AlphaFold

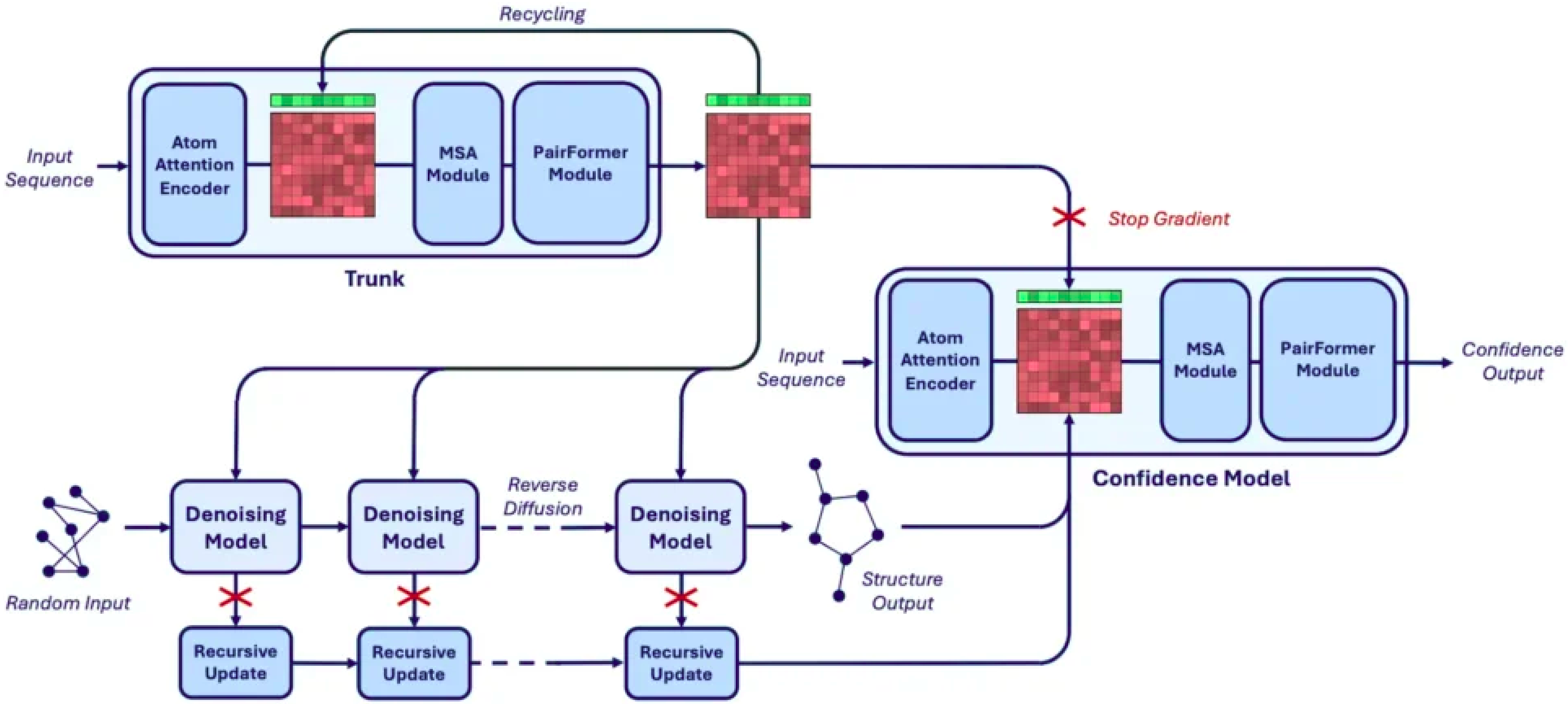
RoseTTAFold

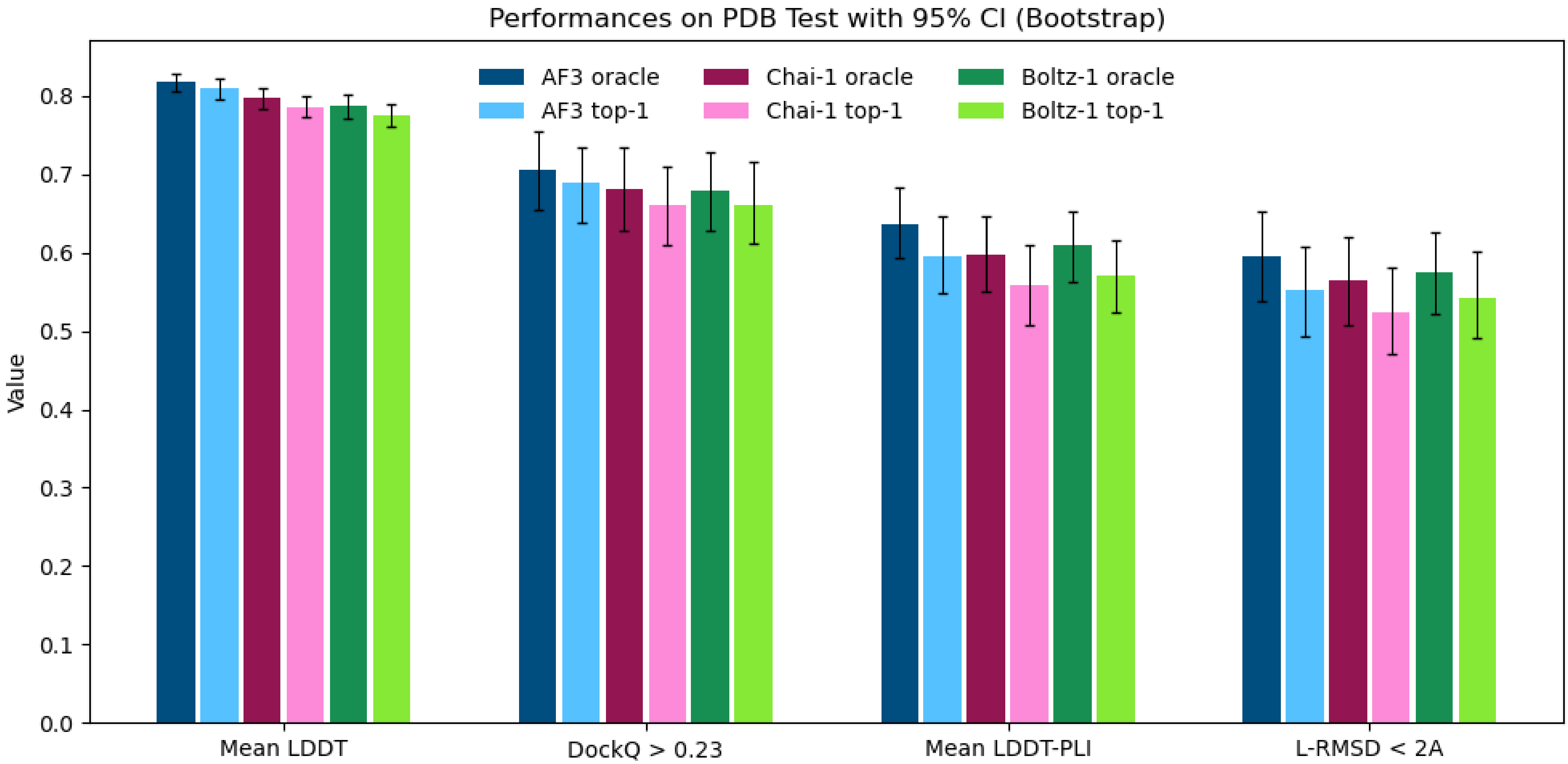
Fold proteins with Boltz-1

HeliXonProtein/
**OmegaFold**

**Protenix: Protein + X**

# Boltz-1


Recycling
Input Sequence
Atom Attention Encoder
MSA Module
PairFormer Module
Trunk
Stop Gradient
Input Sequence
Atom Attention Encoder
MSA Module
PairFormer Module
Confidence Output
Confidence Model
Random Input
Denoising Model
Denoising Model
Reverse Diffusion
Denoising Model
Structure Output
Recursive Update
Recursive Update
Recursive Update


Performances on PDB Test with 95% CI (Bootstrap)
AF3 oracle
AF3 top-1
Chai-1 oracle
Chai-1 top-1
Boltz-1 oracle
Boltz-1 top-1
Value
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Mean LDDT
DockQ > 0.23
Mean LDDT-PLI
L-RMSD < 2A

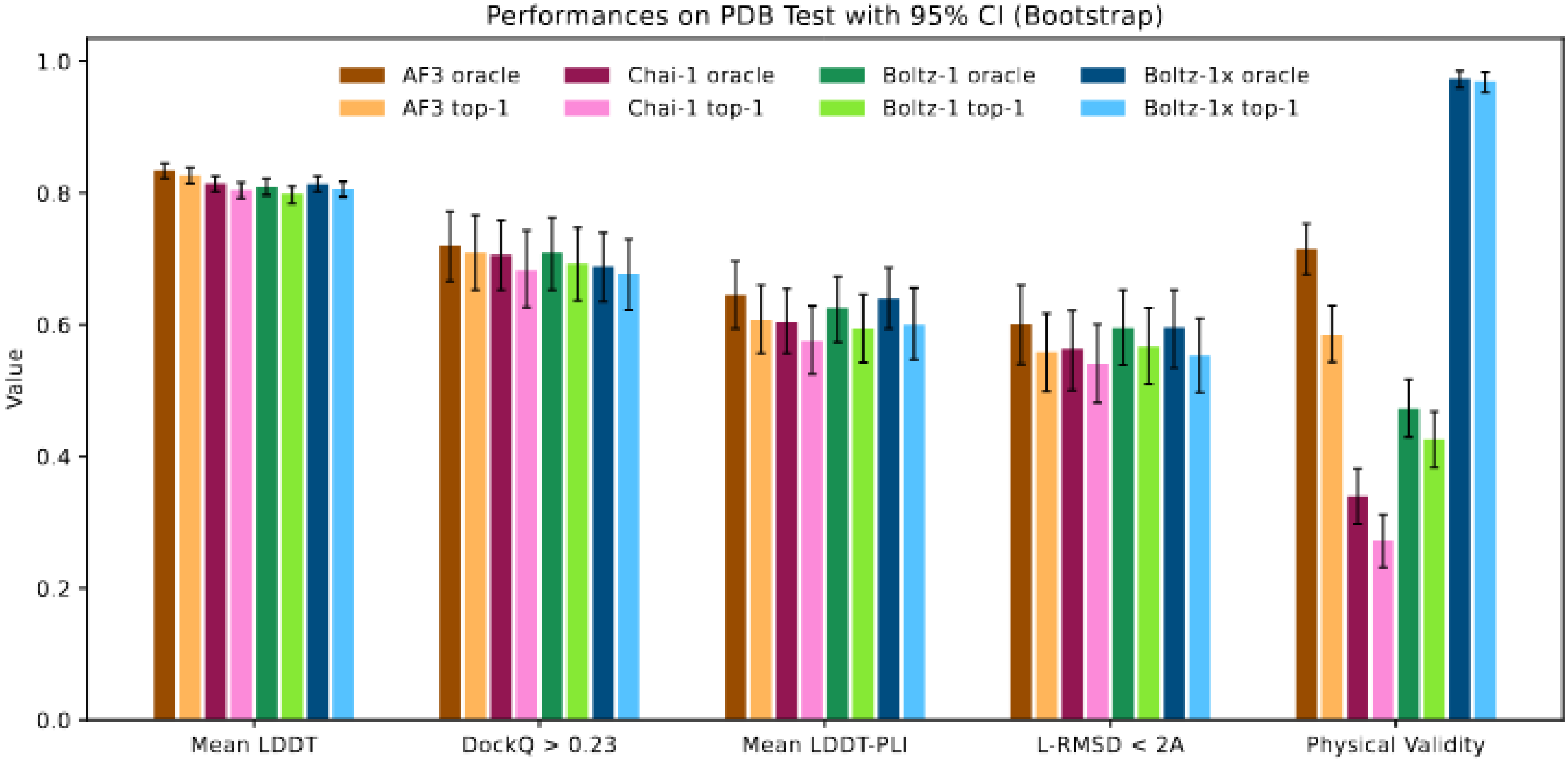


Figure 5: Visual summary of the performance of AlphaFold3, Chai-1, Boltz-1 and Boltz-1x on the test set.

# AlphaFold – like programs

**DeepFold** https://pubmed.ncbi.nlm.nih.gov/36112717/
**RGN2** https://www.nature.com/articles/s41587-022-01432-w
**ProtGPT2** https://www.nature.com/articles/s41467-022-32007-7

https://github.com/RosettaCommons**/RoseTTAFold**

**equifold**
https://www.biorxiv.org/content/10.1101/2022.10.07.511322v1

**DMPfold**
https://github.com/psipred/DMPfold2
https://www.pnas.org/doi/10.1073/pnas.2113348119

**ESMFold**
https://www.biorxiv.org/content/10.1101/2022.07.20.500902v1.abstract
https://github.com/facebookresearch/esm
https://www.nature.com/articles/d41586-022-03539-1
esmatlas.com

**omegafold**
https://www.biorxiv.org/content/10.1101/2022.07.21.500999v1.abstract

**HelixFold**
https://arxiv.org/pdf/2207.05477.pdf

**ProteinBERT**
https://www.biorxiv.org/content/10.1101/2021.05.24.445464v1
http://dx.doi.org/10.1093/bioinformatics/btac020

**trRosettaX-Single** https://doi.org/10.1038/s43588-022-00373-3
https://yanglab.nankai.edu.cn/trRosetta/benchmark_single/

https://analyticsindiamag.com/protein-wars-its-esmfold-vs-alphafold/

# AlphaFold – like programs

#protein seq from backbone
**ProteinMPNN** paper: https://t.co/BLPg2XdmYE
https://colab.research.google.com/github/sokrypton/ColabDesign/blob/v1.1.0/mpnn/examples/proteinmpnn_in_jax.ipynb#scrollTo=GjdIxO4j-Hnn

**ProGen2**: Exploring the Boundaries of Protein Language Models https://arxiv.org/pdf/2206.13517.pdf

**RITA**: a Study on Scaling Up Generative Protein Sequence Models https://arxiv.org/pdf/2205.05789.pdf
https://github.com/lightonai/RITA

**ProT-VAE**: Protein Transformer Variational AutoEncoder for Functional Protein Design
https://www.biorxiv.org/content/10.1101/2023.01.23.525232v1

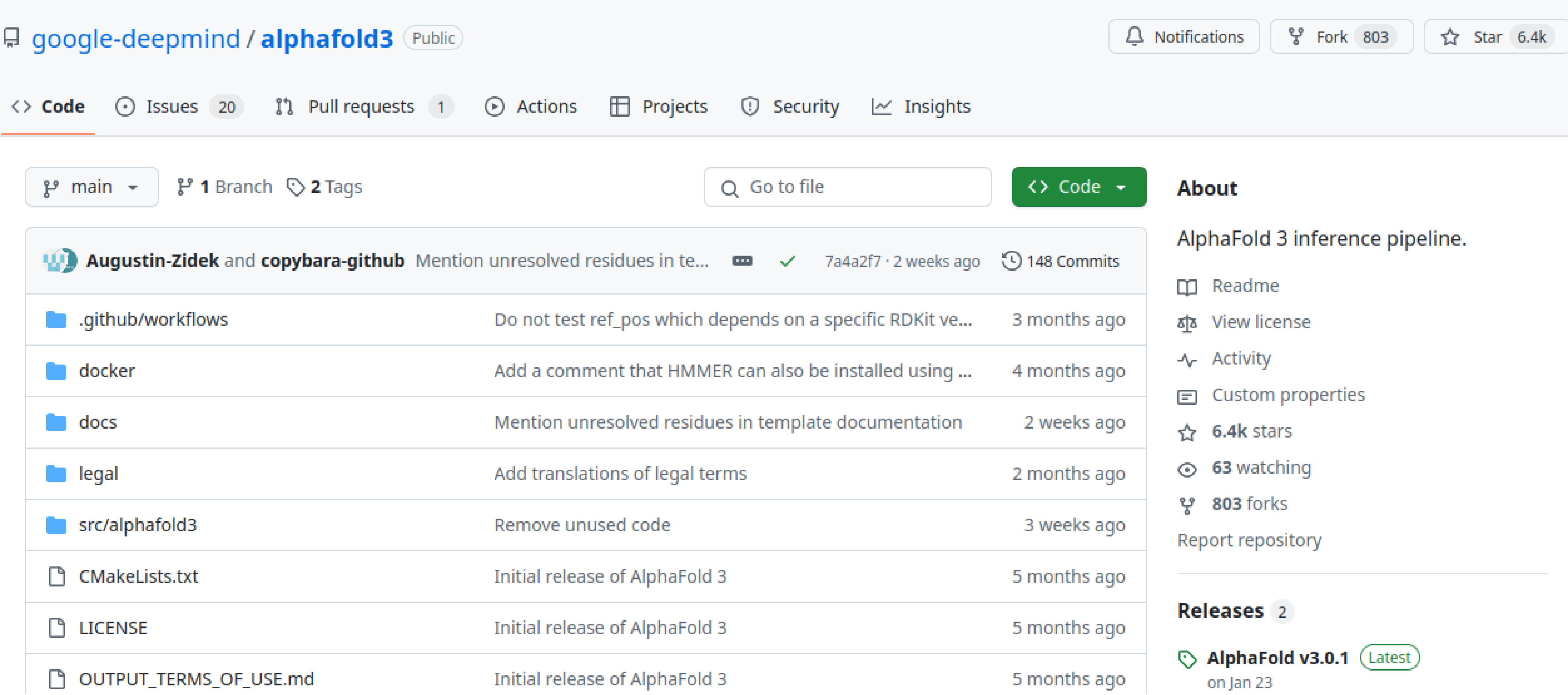
**RSA**
Retrieved Sequence Augmentation for Protein Representation Learning
https://www.biorxiv.org/content/10.1101/2023.02.22.529597v2.abstract
https://github.com/HKUNLP/RSA

**Uni-Fold**
https://github.com/dptech-corp/Uni-Fold#download-from-volcengine
https://colab.research.google.com/github/dptech-corp/Uni-Fold/blob/main/notebooks/unifold.ipynb

**AlphaLink**
https://www.nature.com/articles/s41587-023-01704-zProtein structure prediction with in-cell photo-crosslinking mass spectrometry and deep learning

**EigenFold** Generative Protein Structure Prediction with Diffusion Models
https://arxiv.org/abs/2304.02198
https://github.com/bjing2016/EigenFold

# Scientific Competitions



google-deepmind / **alphafold3** Public

Notifications | Fork 803 | Star 6.4k

**Code** | Issues 20 | Pull requests 1 | Actions | Projects | Security | Insights

main | **1** Branch | **2** Tags | Go to file | Code

**Augustin-Zidek** and **copybara-github** Mention unresolved residues in te... ✓ 7a4a2f7 · 2 weeks ago | 148 Commits

| | | |
|---|---|---|
| .github/workflows | Do not test ref_pos which depends on a specific RDKit ve... | 3 months ago |
| docker | Add a comment that HMMER can also be installed using ... | 4 months ago |
| docs | Mention unresolved residues in template documentation | 2 weeks ago |
| legal | Add translations of legal terms | 2 months ago |
| src/alphafold3 | Remove unused code | 3 weeks ago |
| CMakeLists.txt | Initial release of AlphaFold 3 | 5 months ago |
| LICENSE | Initial release of AlphaFold 3 | 5 months ago |
| OUTPUT_TERMS_OF_USE.md | Initial release of AlphaFold 3 | 5 months ago |

**About**

AlphaFold 3 inference pipeline.

- Readme
- View license
- Activity
- Custom properties
- **6.4k** stars
- **63** watching
- **803** forks

Report repository

**Releases** 2

**AlphaFold v3.0.1** Latest
on Jan 23

alphafoldserver.com

# Scientific Competitions

Critical Assessment of Techniques for Protein Structure Prediction (CASP)

Collective experiment for blind RNA structure prediction (RNA-Puzzles)

Critical Assessment of Prediction of Interactions (CAPRI)

Critical Assessment of Functional Annotation (CAFA)

Critical Assessment of Microarray Data Analysis (CAMDA)

Genome Annotation Assessment Project (GASP)

Bone X-Ray Deep Learning Competition

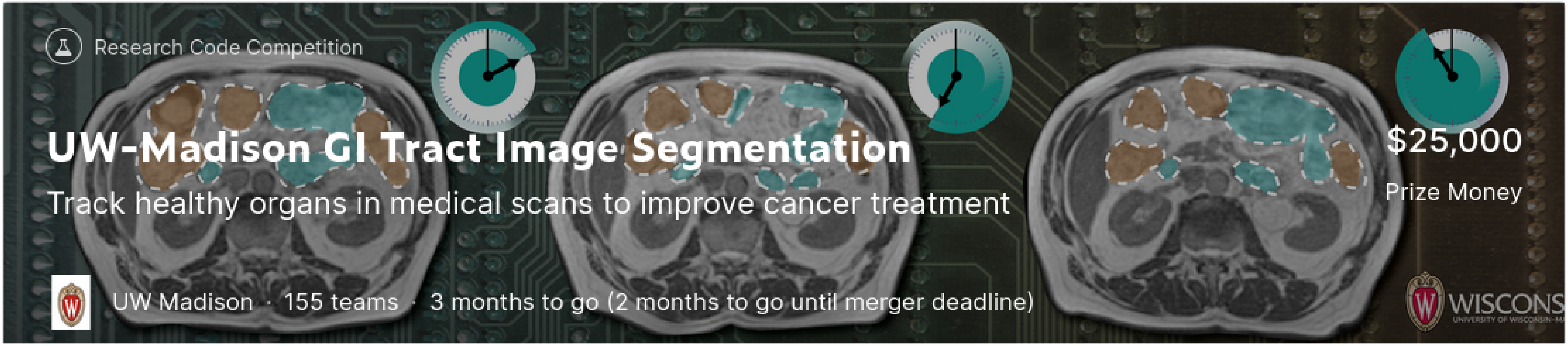
LUng Nodule Analysis 2016

# Scientific Competitions

kaggle

## Competitions

Grow your data science skills by competing in our exciting competitions. Find help in the documentation or learn about Community Competitions.



Research Code Competition

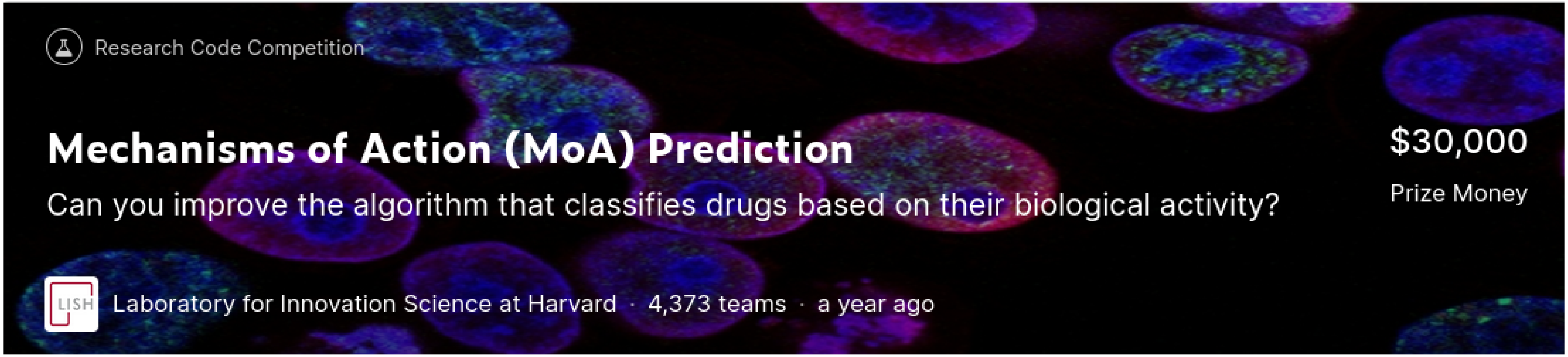
### UW-Madison GI Tract Image Segmentation

Track healthy organs in medical scans to improve cancer treatment

$25,000

Prize Money

UW Madison · 155 teams · 3 months to go (2 months to go until merger deadline)



Research Code Competition

### Mechanisms of Action (MoA) Prediction

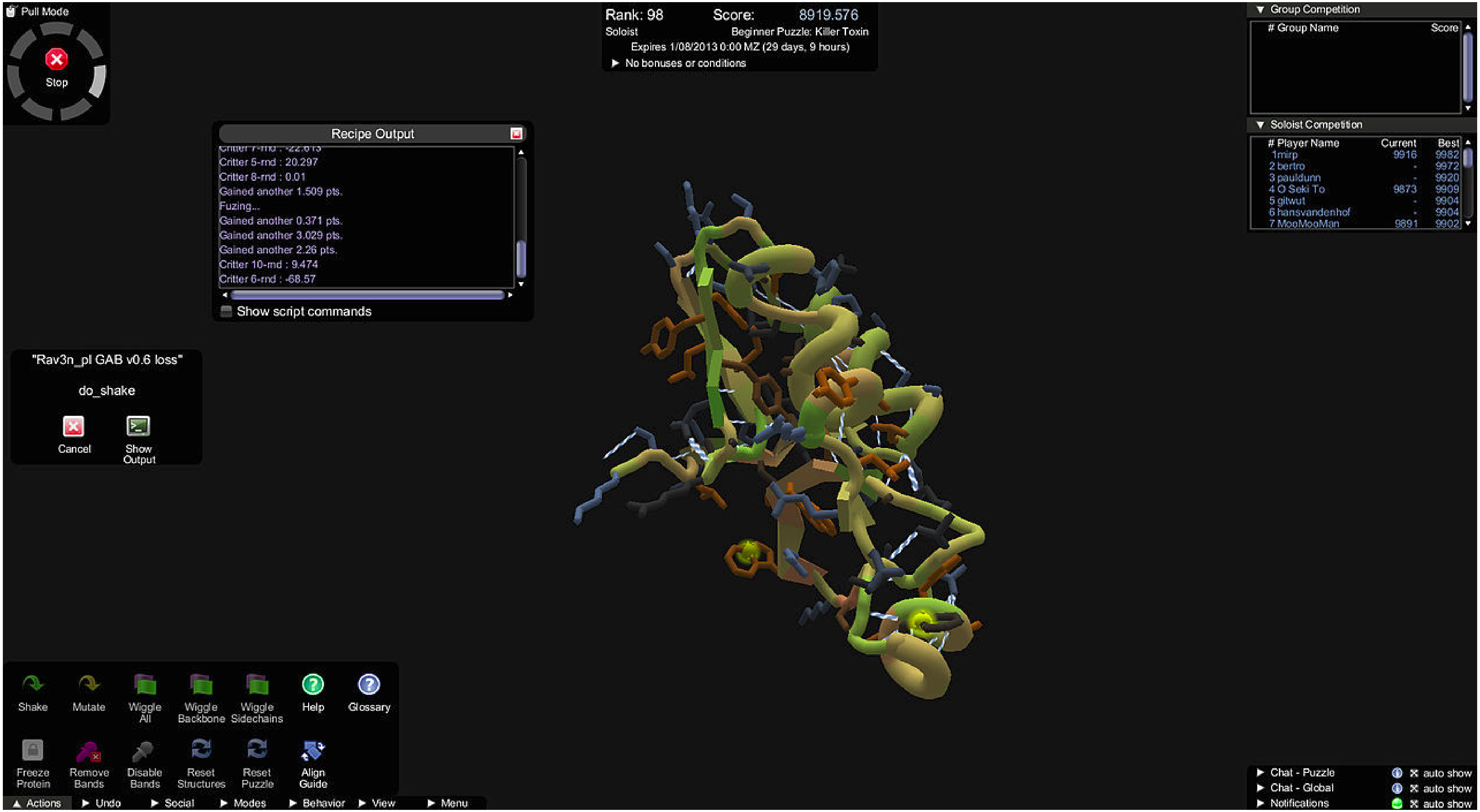
Can you improve the algorithm that classifies drugs based on their biological activity?

$30,000

Prize Money

Laboratory for Innovation Science at Harvard · 4,373 teams · a year ago

# FoldIt - online puzzle video game about protein folding



Rank: 98 Score: 8919.576
Soloist Beginner Puzzle: Killer Toxin
Expires 1/08/2013 0:00 MZ (29 days, 9 hours)
▶ No bonuses or conditions

Recipe Output

```
Critter 7-rnd : -22.613
Critter 5-rnd : 20.297
Critter 8-rnd : 0.01
Gained another 1.509 pts.
Fuzing...
Gained another 0.371 pts.
Gained another 3.029 pts.
Gained another 2.26 pts.
Critter 10-rnd : 9.474
Critter 6-rnd : -68.57
```

□ Show script commands

"Rav3n_pl GAB v0.6 loss"

do_shake

▼ Group Competition

| # | Group Name | Score |
|---|---|---|

▼ Soloist Competition

| # | Player Name | Current | Best |
|---|---|---|---|
| 1 | mirp | 9916 | 9982 |
| 2 | bertro | - | 9972 |
| 3 | pauldunn | - | 9920 |
| 4 | O Seki To | 9873 | 9909 |
| 5 | gitwut | - | 9904 |
| 6 | hansvandenhof | - | 9904 |
| 7 | MooMooMan | 9891 | 9902 |

# Scientific Competitions

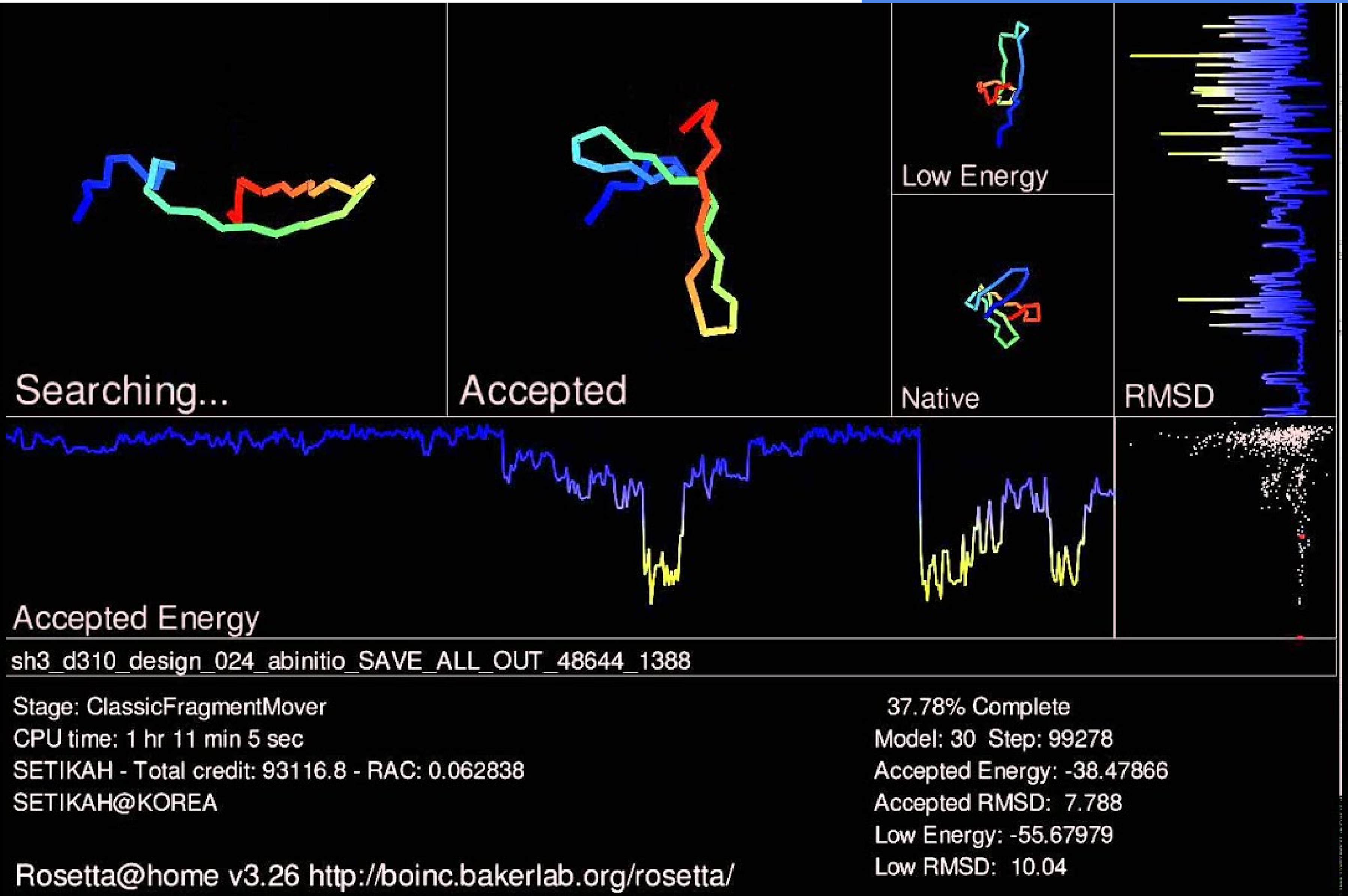
Rosetta@home

Protein Folding, Design, and Docking

You don't have to be a scientist to do science.

By simply running a free program, you can help advance research in medicine, clean energy, and materials science.

Join Rosetta@home



sh3_d310_design_024_abinitio_SAVE_ALL_OUT_48644_1388

Stage: ClassicFragmentMover
CPU time: 1 hr 11 min 5 sec
SETIKAH - Total credit: 93116.8 - RAC: 0.062838
SETIKAH@KOREA

37.78% Complete
Model: 30 Step: 99278
Accepted Energy: -38.47866
Accepted RMSD: 7.788
Low Energy: -55.67979
Low RMSD: 10.04

Rosetta@home v3.26 http://boinc.bakerlab.org/rosetta/

Thank you for your time
and
See you at the next lecture

Any other
questions & comments

**lukaskoz@mimuw.edu.pl**