

Sources of data in experimental structural biology: CryoEM, X-ray crystallography

Saulius Gražulis

Vilnius, 2024

Vilnius University Institute of Biotechnology



Id: slides.tex 11390 2024-06-04 09:28:37Z saulius June 4, 2024



Why databases?

How many crystal structures have been published each year? [Search the COD database:](#)

Why databases?

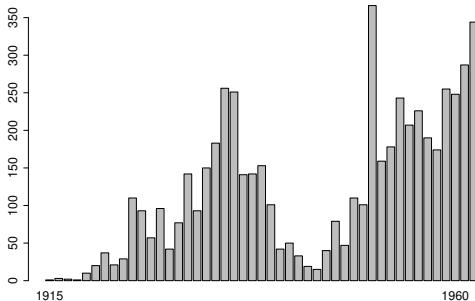
How many crystal structures have been published each year? [Search the COD database:](#)

```
SELECT count(*) AS nr, year FROM data
WHERE year IS NOT NULL AND
GROUP BY year ORDER BY year DESC
```

Why databases?

How many crystal structures have been published each year? [Search the COD database:](#)

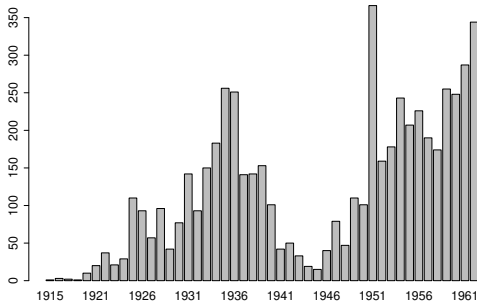
```
SELECT count(*) AS nr, year FROM data
WHERE year IS NOT NULL AND
GROUP BY year ORDER BY year DESC
```



Why databases?

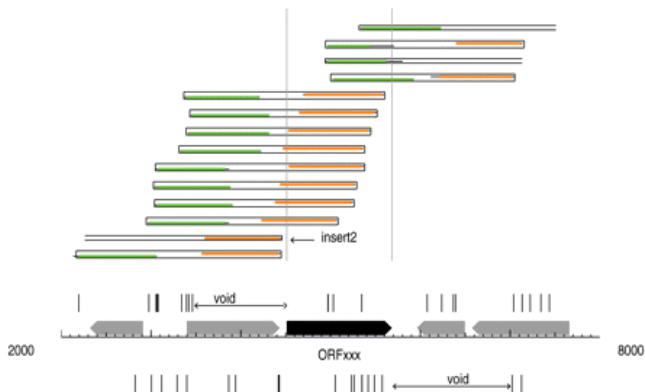
How many crystal structures have been published each year? [Search the COD database:](#)

```
SELECT count(*) AS nr, year FROM data
WHERE year IS NOT NULL AND
GROUP BY year ORDER BY year DESC
```



Discoveries in „raw“ data

Zheng from the NEB team lead by Roberts used the raw sequencing data to discover *active* restriction endonucleases (Zheng et al. 2008):



Importance of data

A group from China made a discovery biochemical pathways for drug addiction without even making and experiment (Li et al. 2008):

The screenshot shows the PLOS Computational Biology article page for "Genes and (Common) Pathways Underlying Drug Addiction" by Chuan-Yun Li, Xizeng Mao, and Liping Wei. The page includes a navigation bar with "Browse", "Publish", and "About" links, a search bar, and a "RESEARCH ARTICLE" label. On the right, there are statistics: 159 Saves, 93 Citations, 52,365 Views, and 2 Shares. Below the article title is a navigation menu with "Article", "Authors", "Metrics", "Comments", and "Related Content". The "Abstract" section is visible, starting with "Drug addiction is a serious worldwide problem with strong genetic and environmental influences. Different technologies have revealed a variety of genes and pathways underlying addiction; however, each individual technology can be biased and incomplete. We integrated 2,243 items of evidence from peer-reviewed publications between 1978 and 2008." On the right side, there are buttons for "Download PDF", "Print", and "Share", along with a "Check for updates" button and a "Subject Areas" section with "Addiction" selected.

PLOS COMPUTATIONAL BIOLOGY

Browse Publish About Search

advanced search

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Genes and (Common) Pathways Underlying Drug Addiction

Chuan-Yun Li, Xizeng Mao, Liping Wei

Published: January 4, 2008 • <http://dx.doi.org/10.1371/journal.pcbi.0040002>

159 Save 93 Citation

52,365 View 2 Share

Article Authors Metrics Comments Related Content

Abstract

Author Summary Introduction Results Discussion

Abstract

Drug addiction is a serious worldwide problem with strong genetic and environmental influences. Different technologies have revealed a variety of genes and pathways underlying addiction; however, each individual technology can be biased and incomplete. We integrated 2,243 items of evidence from peer-reviewed publications between 1978 and 2008.

Download PDF Print Share

Check for updates

Subject Areas

Addiction

<http://slidegur.com/doc/3077570/introducing-bioinformatics-databases>

Amounts of data

High Energy Physics

- LHC generates 1 Terabyte per second.
<http://blogs.discovermagazine.com/cosmicvariance/2006/09/27/lhc-factoids/>
2009-12-07
- „When it starts in 2007 the LHC will ... produce 15 million Gigabytes of data a year“ (15 PB/year $\approx 15 \times 10^{15}$ bytes/year – S.G.)
<http://www.physorg.com/news10895.html>
2009-12-07

Amounts of data

Sequences

- „Release 57.11 of 24-Nov-09 of UniProtKB/Swiss-Prot contains 512994 sequence entries, comprising 180531504 amino acids abstracted from 184920 references.“ (1.81×10^8 a.r., 5.13×10^5 sequences – S.G.)
<http://www.expasy.ch/sprot/relnotes/relstat.html>
2009-12-07
- „As of Tuesday Dec 01, 2009 at 4 PM PST there are 61808 Structures“
<http://www.rcsb.org/pdb/statistics/holdings.do>
2009-12-07

Amounts of data

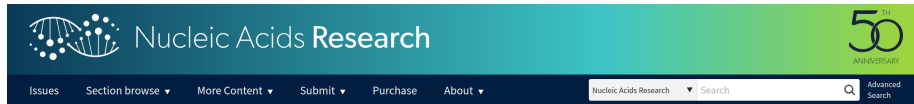
Crystallography

- „Cambridge Structural Database 1 January 2010 Total No. of structures 501857“
<http://www.ccdc.cam.ac.uk/products/csd/statistics/>
2010-12-20
- „Currently there are 126492 entries in the COD“
<http://www.crystallography.net/>
2010-12-20
- „Currently there are 509531 entry in the COD“
<http://www.crystallography.net/>
2024-01-03

Database journals

The Nucleic Acids Research (NAR) Database issue (once per year):

<https://doi.org/10.1093/nar/gkad1173> (Editorial)



The banner features the journal title "Nucleic Acids Research" in white text on a teal background. To the left is a stylized DNA double helix icon. To the right is a green box celebrating the "50TH ANNIVERSARY". Below the title is a dark navigation bar with links: Issues, Section browse, More Content, Submit, Purchase, About. On the right side of the navigation bar is a search box containing "Nucleic Acids Research" and an "Advanced Search" button.



Volume 52, Issue D1
5 January 2024

Article Contents

Abstract

New and updated databases

NAR online molecular biology
database collection

Acknowledgements




JOURNAL ARTICLE

The 2024 *Nucleic Acids Research* database issue and the online molecular biology database collection

Daniel J Rigden , Xosé M Fernández

Nucleic Acids Research, Volume 52, Issue D1, 5 January 2024, Pages D1–D9, <https://doi.org/10.1093/nar/gkad1173>

Published: 30 November 2023 **Article history** ▼

 PDF  Split View  Cite  Permissions  Share ▼

Abstract

The 2024 *Nucleic Acids Research* database issue contains 180 papers from across biology and neighbouring disciplines. There are 90 papers reporting on new databases and 83 updates from resources previously published in the Issue. Updates from databases most recently published elsewhere account for a



Email alerts

[Article activity alert](#)
[Advance article alerts](#)
[New issue alert](#)
[Subject alert](#)

<https://academic.oup.com/nar/article/52/D1/D1/7456037#434765858>

PDF Split View Cite Permissions Share

Nucleic Acids Research

Search



Volume 52, Issue D1
5 January 2024

Article Contents

Abstract

New and updated databases

NAR online molecular biology database collection

Acknowledgements

Funding

References

Comments (0)

Table 1. Descriptions of new databases in the 2023 NAR Database issue

Database name	URL	Short description
A3D Model Organism Database	http://biocomp.chem.uw.edu.pl/A3D2/MODB	Aggregation predictions for model organism proteomes
ABC-HuMi	https://www.ccb.uni-saarland.de/abc_humi/	Biosynthetic Gene Clusters in the Human Microbiome
ADC-DB	https://idrlab.org/adcdb/	Antibody-drug conjugates
AgeAnnoMO	https://relab.xidian.edu.cn/AgeAnnoMO/#/	Multi-omics of animal ageing
AGIDB	https://agidb.pro/	Animal Genotype Imputation DataBase
AnimalMetaOmics	https://yanglab.hzau.edu.cn/animalmetaomics#	Multi-omics of animal microbiomes
ATLAS	https://www.dsimb.inserm.fr/ATLAS	ATLAS of protein dynamics
BioExcel-CV19	https://bioexcel-cv19.bsc.es/#/	Molecular Dynamics of Covid-19 proteins
BioKA	https://ngdc.cncb.ac.cn/bioka	Biomarker Knowledgebase for Animals

Citing articles via

Google Scholar

Latest Most Read Most Cited

The loss of DNA polymerase epsilon accessory subunits POLE3–POLE4 leads to BRCA1-independent PARP inhibitor sensitivity

The fission yeast ortholog of Coilin, Mug174, forms Cajal body-like nuclear condensates and is essential for cellular quiescence

DMC1 and RAD51 bind FxxA and FxPP motifs of BRCA2 via two separate interfaces

The translocation activity of Rad54 reduces crossover outcomes during homologous recombination

Unlocking the potential of protein-derived peptides to target G-quadruplex DNA: from recognition to anticancer activity

<https://www.oxfordjournals.org/nar/database/c/>



You are here: [NAR Journal Home](#) » Database Summary Paper Categories

NAR Database Summary Paper Category List

[Nucleotide Sequence Databases](#)

[RNA sequence databases](#)

[Protein sequence databases](#)

[Structure Databases](#)

[Genomics Databases \(non-vertebrate\)](#)

[Metabolic and Signaling Pathways](#)

[Human and other Vertebrate Genomes](#)

[Human Genes and Diseases](#)

[Microarray Data and other Gene Expression Databases](#)

[Proteomics Resources](#)

[Other Molecular Biology Databases](#)

[Organelle databases](#)

[Plant databases](#)

[Immunological databases](#)

[Cell biology](#)

- ▶ [Compilation Paper](#)
- ▶ [Category List](#)
- ▶ [Alphabetical List](#)
- ▶ [Category/Paper List](#)
- ▶ [Search Summary Papers](#)

Database journals: Database

<https://academic.oup.com/database>

OXFORD ACADEMIC Journals Books   Sign in through your institution

DATABASE

The Journal of Biological Databases and Curation



Volumes ▾ Submit ▾ Alerts About ▾ Database Search  Advanced Search



Latest Volume
Volume 2024
2024

Impact Factor
5.8

5 year Impact Factor
4.6

Editor-in-Chief
David Landsman

About the journal

Database: The Journal of Biological Databases and Curation provides an open access platform for the presentation of novel ideas in database research and biocuration, and aims to help strengthen the bridge between ...

Find out more

Data journals: Scientific Data

<https://www.nature.com/sdata/>

scientific **data**

[View all journals](#)

[Search](#)

[Log in](#)

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾

[Sign up for alerts](#) 🔔

[RSS feed](#)

[nature](#) > [scientific data](#)

Two decades of fumigation data from the Soybean Free Air Concentration Enrichment facility

Elise Kole Aspray, Timothy A. Mies ... Elizabeth A. Ainsworth
[Data Descriptor](#) | 20 April 2023



<https://www.mdpi.com/journal/data>



Journals

Topics

Information

Author Services

Initiatives

About

Sign In / Sign Up

Submit

Search for Articles:

Title / Keyword

Author / Affiliation / Email

Data

All Article Types

Search

Advanced

Journals / Data



data

Submit to *Data*

Review for *Data*

Data

Data is a peer-reviewed, open access journal on data in science, with the aim of enhancing data transparency and reusability. The journal publishes in two sections: a section on the collection, treatment and analysis methods of data in science; a section publishing descriptions of scientific and scholarly datasets (one dataset per paper). The journal is published monthly online by MDPI.

IMPACT
FACTOR
2.6

CITESCORE
4.6

E-Mail Alert

Add your e-mail address to receive forthcoming issues of this journal:

Enter Your E-Mail Address...

Subscribe

Scientific data sources

1 Generic data archives (Data Dryad, FigShare, MIDAS, Zenodo ...);



2



① Generic data archives (Data Dryad, FigShare, MIDAS, Zenodo ...);

- Zenodo <https://doi.org/10.5281/zenodo.3560693>
- Zenodo <https://zenodo.org/record/3841841>

②

-
-
-
-

① Generic data archives (Data Dryad, FigShare, MIDAS, Zenodo ...);

- Zenodo <https://doi.org/10.5281/zenodo.3560693>
– poster (PDF)...
- Zenodo <https://zenodo.org/record/3841841>

②

-
-
-
-

① Generic data archives (Data Dryad, FigShare, MIDAS, Zenodo ...);

- Zenodo <https://doi.org/10.5281/zenodo.3560693>
– poster (PDF)...
- Zenodo <https://zenodo.org/record/3841841>
– linked COVID-19 data (Turtle .ttl)...

②

-
-
-
-

Scientific data sources

- 1 Generic data archives (Data Dryad, FigShare, MIDAS, Zenodo ...);
 - Zenodo <https://doi.org/10.5281/zenodo.3560693>
– poster (PDF)...
 - Zenodo <https://zenodo.org/record/3841841>
– linked COVID-19 data (Turtle .ttl)...
- 2 Specialised data archives (PDB, COD, NCBI, SwissProt, EuropePMC, PubMed (!));
 -
 -
 -
 -

Scientific data sources

- 1 Generic data archives (Data Dryad, FigShare, MIDAS, Zenodo ...);
 - Zenodo <https://doi.org/10.5281/zenodo.3560693>
– poster (PDF)...
 - Zenodo <https://zenodo.org/record/3841841>
– linked COVID-19 data (Turtle .ttl)...
- 2 Specialised data archives (PDB, COD, NCBI, SwissProt, EuropePMC, PubMed (!));
 - <https://www.crystallography.net/cod/1557684.cif>
 - <https://www.crystallography.net/cod/1544162.html>
 - <https://www.pdb.org/pdb/files/1KNV.cif>
 - <https://www.rcsb.org/structure/2IXS>

The Protein Data Bank

Three major repositories in different continents, governed by the wwPDB consortium:

<https://www.wwpdb.org/>

WORLDWIDE PDB PROTEIN DATA BANK VALIDATION • DEPOSITION • DICTIONARIES • DOCUMENTATION • TASK FORCES • FTP • STATISTICS • ABOUT • wwPDB Consortium

Since 1971, the Protein Data Bank archive (PDB) has served as the single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies.

The Worldwide PDB (wwPDB) organization manages the PDB archive and ensures that the PDB is freely and publicly available to the global community.

Celebrating 50 Years of the PDB

Validate Structure
or View validation reports

Deposit Structure
All Deposition Resources

Download Archive
Instructions

Vision and Mission

Vision
Sustain freely accessible, interoperating Core Archives of structure data and metadata for biological macromolecules as an enduring public good to promote basic and applied research and education across the sciences.

Mission

- Manage the wwPDB Core Archives as a public good according to the FAIR Principles.
- Provide expert deposition, validation, biocuration, and remediation services at no charge to Data Depositors worldwide.
- Ensure universal open access to public domain structural biology data with no limitations on usage.
- Develop and promote community-endorsed data standards for archiving and exchange of global structural biology data.

wwPDB Members

Protein Data Bank Japan **PDBj**
Supports browsing in multiple languages such as Japanese, Chinese, and Korean; SeSAW identifies

wwPDB Resources

Data Dictionaries

- Macromolecular Dictionary (PDB/mmCIF)
- Small Molecule Dictionary (CCD)
- Peptide-like antibiotic and inhibitor molecules (BIRD)

Biocuration

- Procedures and policies
- Improvements for consistency and accuracy

Community Input:
Task Forces and Working Groups

- Validation Task Forces (X-ray, NMR, 3DEM)
- Small Angle Scattering Task Force
- PDB-mmCIF Working Group
- Hybrid Integrative Methods Task Force
- Ligand Validation Workshop

PDB Data Growth & Usage Statistics

- Depositions: by data center, by year, and by depositor location
- Downloads: by year for all entries

News & Announcements

11/02/2021
November Workshops on Open-Source Tools for Chemistry

Join the Royal Society of Chemistry for two webinars on Protein Data Bank at 50: Accessing, Understanding, and Assessing PDB Data

[Read more](#)

10/27/2021
Obituary for John Westbrook

John D. Westbrook Jr. (1967-2021) passed away on October 18, 2021. He was incredibly beloved and respected by his colleagues at Rutgers and throughout the world, known for his dry wit and endless enthusiasm for thinking about all aspects of

The Protein Data Bank: RCSB PDB

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

The screenshot shows the RCSB PDB website homepage. At the top, there is a navigation bar with links for Deposit, Search, Visualize, Analyze, Download, Learn, More, Documentation, and Careers. A search bar is located on the right side of the navigation bar. Below the navigation bar, the RCSB PDB logo is displayed, along with the text "184202 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education". A search bar is also present below the logo, with the text "Enter search terms or PDB ID(s)". Below the search bar, there are several logos for partner organizations, including PDB-101, PDBe, EMBL Data Resource Centre, and the Worldwide Protein Data Bank Foundation. A banner for "Celebrating 50 Years of Protein Data Bank" is also visible. Below the banner, there is a section for "Developers: Join the RCSB PDB Team" and "Explore Open Positions". The main content area is divided into several sections. On the left, there is a sidebar with navigation links: Welcome, Deposit, Search, Visualize, Analyze, Download, and Learn. The main content area features a section titled "A Structural View of Biology" with a description of the resource and its purpose. Below this section, there are two images: "COVID-19 CORONAVIRUS Resources" and "Celebrating Protein Data Bank". To the right of the main content area, there is a section titled "November Molecule of the Month" featuring a 3D structure of Acetohydroxyacid Synthase. At the bottom of the page, there is a navigation bar with links for Latest Entries, Features & Highlights, News, and Publications.

RCSB PDB Deposit Search Visualize Analyze Download Learn More Documentation Careers MyPDB

RCSB PDB 184202 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

Enter search terms or PDB ID(s). Help

Advanced Search | Browse Annotations

Celebrating 50 YEARS OF Protein Data Bank

Developers: Join the RCSB PDB Team Explore Open Positions

Welcome

Deposit

Search

Visualize

Analyze

Download

Learn

A Structural View of Biology

This resource is powered by the Protein Data Bank archive—information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

COVID-19 CORONAVIRUS Resources

Celebrating PROTEIN DATA BANK

November Molecule of the Month



Acetohydroxyacid Synthase

Latest Entries All of Nov 16 2021 Features & Highlights News Publications

Contact Us

4

The Protein Data Bank: PDBe

<https://www.ebi.ac.uk/pdbe/>

The screenshot shows the PDBe website interface. At the top, there is a navigation bar with 'EMBL-EBI' logo and 'Protein Data Bank in Europe' title. Below this is a search bar with a 'Search' button and a 'Advanced search' link. A secondary navigation bar contains links for 'PDBe home', 'Deposition', 'PDBe services', 'PDBe training', 'Documentation', 'About PDBe', and 'COVID-19'. The main content area is divided into several sections: 'New PDBe-KB COVID-19 Data Portal' with a sub-section for 'COVID-19 Data Portal' and a link to visit the portal; 'Featured structure' with a cat image and text about microbial pathogens; 'News' with articles on '3D-Beacons Network: protein structure data, all in one place' and 'AlphaFold's protein structure predictions now available to explore'; and 'Events' with a message that no upcoming events are found. On the right side, there is a 'Popular' list of resources, 'Latest archive statistics' showing 184202 entries as of 17 November 2021, and a 'Tweets by @PDBeurope' section featuring a tweet about Alcohol dehydrogenases (ADH) enzymes. The bottom of the page shows a video player with a play button over a protein structure visualization.

The Protein Data Bank: PDBj

<https://pdbj.org/>

The screenshot displays the PDBj website homepage. At the top, there is a navigation bar with the PDBj logo and the text "Protein Data Bank Japan". A search bar is located in the top right corner. Below the navigation bar, there are several main sections:

- Home:** Includes links for Top Page, Statistics, Help, FAQ, Contact Us, Cite Us / Terms and Conditions, and Links.
- About PDBj:** A paragraph describing the project team operating under the Joint Usage and Research activities of the Institute for Protein Research, Osaka University, in collaboration with RCSB PDB in the USA, EMBL-EBI in the EU, and EMBL in the EU. It also mentions support from JST-NISDC and AMED-BINDS.
- Find the service you need:** A section with a search box and a list of services categorized by PDB, EMDB, search, deposition, viewer, education/dictionary, NMR, electron microscopy, secondary structure, sequence, similarity, function prediction, chemical component, structure prediction, binding site, surface structure, 3D structure, genome, RDF, SPARQL, gene, and drug. There are buttons for "Show all services" and "Reset".
- Latest news:** A list of recent news items with dates and titles, such as "226 new PDB entries have been released on 2021-11-17" and "November 24, 2021 (Wed), 'PDB50 Anniversary Symposium in Asia' will be held online".
- Right sidebar:** Contains several featured items: "7616 Latest new entries" with a colorful protein structure, "Molecule of the Month" featuring "263: Acetylhydroxyacid Synthase" with a blue protein structure and an "Article List" link, "PROTEIN DATA BANK" logo, "Hot Structural News on COVID-19" with a circular protein structure, "EM Navigator" logo, "BMRBj" logo, and "EMPIAR" logo.

The Protein Data Bank: the FTP site

<http://ftp.wwpdb.org/pub/pdb/>[accessed: 2021]

<https://files.wwpdb.org/pub/pdb/>[accessed: 2024-01-03T07:56+00:00]

PDB - FTP Archive over HTTP

Name	Last modified	Size
« Parent Directory		-
■ compatible/	2020-07-04 05:35	-
■ data/	2020-07-04 05:35	-
■ derived_data/	2016-03-29 05:58	-
■ doc/	2011-07-08 11:23	-
■ holdings/	2021-09-10 11:41	-
■ refdata/	2021-09-10 06:06	-
■ software/	2014-05-09 07:35	-
■ validation_reports/	2021-11-19 07:16	-
📄 advisory.txt	2008-04-14 07:43	2.1K
📄 ls-lR	2021-11-19 14:46	340M
📖 README	2016-03-16 06:54	1.1K
📄 welcome.msg	2007-06-29 06:45	1.0K

For more information about PDB file downloads please see the wwPDB website.

PDB contains macromolecular structures solved by:

- X-ray crystallography
- CryoEM
- NMR

The AlphaFold prediction database

<https://deepmind.com/research/open-source/alphafold-protein-structure-database>

DeepMind > Research > AlphaFold Protein Structure Database

AlphaFold Protein Structure Database

AlphaFold is our AI system that predicts a protein's 3D structure from its amino acid sequence. In **CASP14**, AlphaFold was the top-ranked protein structure prediction method by a large margin, producing predictions with high accuracy, many of which are competitive with experimentally-determined measurements.

We've partnered with Europe's flagship laboratory for life sciences - EMBL's European Bioinformatics Institute ([EMBL-EBI](#)) - to create the AlphaFold Protein Structure Database to make these predictions freely available to the scientific community.

The initial release of the database covers all of the 20,000 proteins in the human proteome, along with the proteomes of several other biologically significant organisms, from E.coli to yeast, and from the fruit fly to the mouse. In the coming months we plan to expand the database to cover a large proportion of all the 300 million proteins catalogued in the [UniProt database](#).

The AlphaFold Protein Structure Database will continue to expand over time, so if you can't find what you're looking for right now, please follow [DeepMind](#) and [EMBL-EBI](#) social channels for updates. In the meantime, you can use the AlphaFold [source code](#) to predict the structures of proteins not yet in the AlphaFold DB, and the [Colab notebook](#) to run individual sequences.

We would love to hear your feedback and understand how AlphaFold has been useful in your research. Share your stories at alphafold@deepmind.com.

<https://deepmind.com/>

“Our models are trained on structures extracted from the PDB” (Senior et al. 2020).

Raw diffraction data

<https://proteindiffraction.org/>

The screenshot shows the homepage of Proteindiffraction.org. At the top, there is a navigation bar with links for Home, About, Browse, Statistics, Submit data, and News, along with a Login button. A search bar for diffraction images is also present. The main content area features the NIH logo and the title "Integrated Resource for Reproducibility in Macromolecular Crystallography". A paragraph describes the project's funding by the Targeted Software Development award and its goal of developing tools for protein diffraction data. It lists currently indexed projects (5815) and datasets (9246). A section on data usage states that data is freely available under a CC0 license, with a request to credit the source. Below this, a row of icons represents various site functions: Browse & search, Statistics, Submit data, Publications, Citing, Beamlines, and COVID-19 Data. At the bottom, there are search examples and logos for partner organizations like FEBS, Protein Science, and IUCr. The footer includes the text "Created by Minor lab at University of Virginia".

Raw diffraction data

<https://data.sbgrid.org/>

The screenshot shows the SBGrid Data Bank website. At the top, there is a navigation bar with the SBGrid logo, a 'Publication Guidelines' link, and menu items for 'Data', 'About', 'Get Help', and 'For Depositors'. Below the navigation bar, a main heading reads 'The SBGrid Data Bank', followed by a paragraph describing the bank's support for X-ray diffraction, MicroED, and LLSM datasets. Two buttons, 'VIEW DATA' and 'DEPOSIT DATA', are positioned to the right. A dark grey bar below this section displays statistics: '669 Datasets', '91 Institutions', and '581 Structures'. The main content area features three columns: 'Deposit' (with a stack of coins icon), 'Explore' (with a computer monitor icon), and 'Cite' (with a citation icon). Each column contains a brief description of the action. At the bottom, a section titled 'RECENTLY PUBLISHED DATASETS' displays a horizontal carousel of four 3D protein structure models. The carousel includes navigation arrows and a page number '4'.

SBGrid
Data Bank

Publication Guidelines

Data • About • Get Help • For Depositors •

The SBGrid Data Bank

We support publication of X-ray diffraction, MicroED, LLSM datasets, as well as structural models. All visitors can access our Laboratory and Institutional Collections. All structural biologists are invited to deposit datasets.

[VIEW DATA](#) [DEPOSIT DATA](#)

669 Datasets 91 Institutions 581 Structures

Deposit
Share your data with the community. Every dataset deposited with SBDB receives a unique DOI and its own landing page here.

Explore
Browse all published datasets and download via rsync.

Cite
Give credit to the data used in your research. Every dataset published with SBDB generates its own citation to be used within manuscripts.

RECENTLY PUBLISHED DATASETS

4

Small angle scattering database

<https://www.sasbdb.org/>

The screenshot shows the SASBDB (Small Angle Scattering Biological Data Bank) homepage. At the top, there is a navigation bar with links for Home, Browse, Submit data, About SASBDB, and Help. A search bar is located on the right side. The main content area features a section titled "Curated repository for small angle scattering data and models". Below this, there is a featured article for "SASDKY3 – Minimal trans Varkud Satellite (VS) ribozyme in 5mM MgCl2". The article includes a sample description, buffer conditions, and experimental details. It also features a 3D ribbon diagram of the ribozyme structure and a SAXS data plot. Below the featured article, there are several smaller article thumbnails with titles like "LDLA linker region of human P...". At the bottom of the page, there are four pie charts representing the distribution of data by source: Proteins, Microorganisms de France, Model ions, and Publications.

SASBDB
Small Angle Scattering Biological Data Bank

Sign in | Register

Advanced search E.g. PDBID, BSA, Nucleic Acids, PDB

Home Browse Submit data About SASBDB Help

Curated repository for small angle scattering data and models

Small angle scattering (SAXS) of X-ray and neutrons provides structural information on biological macromolecules in solution at a resolution of 1-2 nm.
SASBDB is a fully searchable curated repository of freely accessible and downloadable experimental data, which are deposited together with the relevant experimental conditions, sample details, derived models and their fits to the data.

SASBDB currently contains:
2474 experimental data sets
3578 models
436 experimental data sets on hold
917 models on hold

Recent depositions:

SASDKY3 – Minimal trans Varkud Satellite (VS) ribozyme in 5mM MgCl2

Sample: Neurospora Varkud Satellite minimal trans ribozyme monomer, 33 kDa
Neurospora crassa RNA

Buffer: 50 mM MES, 50 mM KCl, 5 mM MgCl2, pH 6.0

Experiment: SAXS data collected at Kamore BioKohler L with Metajet, Université de Montréal on 2019 Sep 19

Keywords: pI_{center} 3.0 res
 D_{max} 11.0 res
 V_{volume} 44 res³

An integrative NMR-SAXS approach for structural determination of large RNAs defines the substrate-free state of a DMS1-cleaving Neurospora Varkud Satellite ribozyme. *Nucleic Acids Research* (2021), Degraen P, Desjardins G, Legault P

LDLA linker region of human P...
Structural insights into the LDLA...
J Mol Biol (2021) 102171 (2021 Feb 1)

Tau protein, 2NAR isoform
Tau protein in Molecular Biochemis...
Mol Biol (2021) 102171 (2021 Feb 1)

Sensory rhodopsin 8 - trans
Antibodies in and completion...
Acta Crystallographica Section D (2021)

Complex of Archaeoglobus fr...
Prokaryotic Argonase from Arch...
Bio Rep (2021) 102171 (2021 Feb 1)

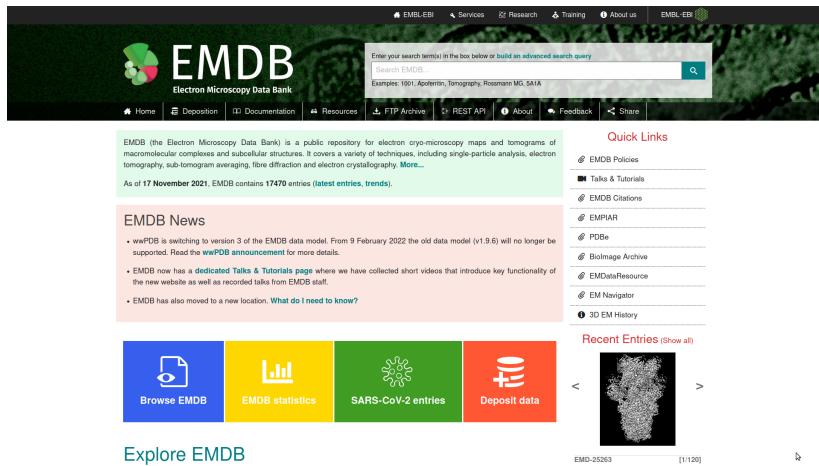
Transcription factors USF1
The USF domain of USF1...
International Journal of Biological...

Browse the contents according to:

- Proteins
- Microorganisms de France
- Model ions
- Publications

Electron microscopy densities

<https://www.ebi.ac.uk/emdb/>



The screenshot shows the EMDB (Electron Microscopy Data Bank) homepage. At the top, there is a navigation bar with links for Home, Deposition, Documentation, Resources, FTP Archive, REST API, About, Feedback, and Share. The main header features the EMDB logo and a search bar with the text "Enter your search term(s) in the box below or build an advanced search query". Below the search bar, there are examples of search terms: "1001, Apopterin, Tomography, Rossmann MG, SARA".

The main content area is divided into several sections:

- EMDB (the Electron Microscopy Data Bank) is a public repository for electron cryo-microscopy maps and tomograms of macromolecular complexes and subcellular structures. It covers a variety of techniques, including single-particle analysis, electron tomography, sub-tomogram averaging, fibre diffraction and electron crystallography. [More...](#)**
- As of 17 November 2021, EMDB contains 17470 entries (latest entries, trends).**
- EMDB News**
 - wwPDB is switching to version 3 of the EMDB data model. From 9 February 2022 the old data model (v1.9.6) will no longer be supported. Read the [wwPDB announcement](#) for more details.
 - EMDB now has a [dedicated Talks & Tutorials page](#) where we have collected short videos that introduce key functionality of the new website as well as recorded talks from EMDB staff.
 - EMDB has also moved to a new location. [What do I need to know?](#)
- Quick Links**
 - EMDB Policies
 - Talks & Tutorials
 - EMDB Citations
 - EMPIAR
 - PDBe
 - BioImage Archive
 - EMDataResource
 - EM Navigator
 - 3D EM History
- Recent Entries (Show all)**
 - EMD-25263 [1/120]

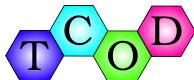
At the bottom, there are four colored buttons: "Browse EMDB" (blue), "EMDB statistics" (yellow), "SARS-CoV-2 entries" (green), and "Deposit data" (orange). Below these buttons is the text "Explore EMDB".

COD and TCOD databases

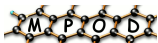
What is COD and TCOD?



<http://www.crystallography.net/cod>
> 480 000 entries



<http://www.crystallography.net/tcod>
> 2900 entries (ready to grow to > 10^7 ?)



<http://mpod.cimav.edu.mx/>
> 300 entries



<http://www.crystallography.net/pcod>
> 10^6 entries (ready to grow to > 10^8 ?)



<http://solsa.crystallography.net/rod/>
> 1100 entries

(Gražulis et al. 2009; Gražulis et al. 2012; Pepponi et al. 2012; Fuentes-Cobas et al. 2017; Mendili et al. 2019)

The COD database

The COD (<https://www.crystallography.net/>) is the largest to-date open access database of *experimental* crystal structures (Gražulis et al. 2009; Gražulis et al. 2012)

- Contains crystal data for compounds:
 - organic
 - metal-organic and organometallic
 - inorganic (including minerals, also from AMCSD)
- experimentally determined structures
- covers most peer-reviewed journals, updated daily
- over **500 000** records at the moment
- curated, both automatically and manually
- versioned

A typical COD record

<http://www.crystallography.net/cod/1559914.html> (deposited 2021-11-21)



Crystallography Open Database

COD Home

[Home](#)
[What's new?](#)

Accessing COD Data

[Browse](#)
[Search](#)
[Search by structural formula](#)

Add Your Data

[Deposit your data](#)
[Manage depositions](#)
[Manage/release prepublications](#)

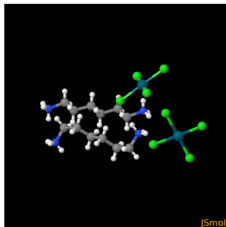
Documentation

[COD Wiki](#)
[Obtaining COD License](#)
[Querying COD](#)
[Citing COD](#)
[COD Mirrors](#)
[Advice to donators](#)
[Useful links](#)

Information card for entry 1559914

[1559913](#) << [1559914](#) >> [1559915](#)

Preview



Chemical name	hexane-1,6-diaminium tetrachloro-palladium(II)
Formula	C6 H18 Cl4 N2 Pd
Calculated formula	C6 H18 Cl4 N2 Pd
Title of publication	Experimental and Theoretical Evidence of Attractive Interactions between Dianions: [PdCl4]2-...[PdCl4]2-
Authors of publication	Zierkiewicz, Wiktor; Michalczyk, Mariusz; Maris, Thierry; Wysokinski, Rafal; Scheiner, Steve
Journal of publication	Chemical Communications
Year of publication	2021
a	7.2281 ± 0.0007 Å
b	8.1281 ± 0.0005 Å
c	11.7212 ± 0.0012 Å

CIF – Crystallographic Interchange Framework

Created and maintained by the International Union of Crystallography, IUCr (Hall et al. 1991).

examples/data/2100858-head.cif:

```
data_2100858
loop_
  _publ_author_name
  'Buttner, R. H.'
  'Maslen, E. N.'
  _publ_section_title
;
  Structural parameters and electron difference density in BaTiO3-
;
  _journal_issue          6
  _journal_name_full     'Acta Crystallographica Section B'
  _journal_page_first    764
  _journal_page_last     769
  _journal_volume        48
  _journal_year          1992
  _chemical_compound_source 'synthetic, from a mixture of KF:KMoO4:BaTiO3'
  _chemical_formula_sum   'Ba O3 Ti'
  _chemical_formula_weight 233.24
  _symmetry_cell_setting  tetragonal
  _symmetry_space_group_name_Hall 'P 4 -2'
  _symmetry_space_group_name_H-M 'P 4 m m'
  _cell_length_a          3.9998(8)
  _cell_length_b          3.9998(8)
  _cell_length_c          4.0180(8)
```

Semantic descriptions

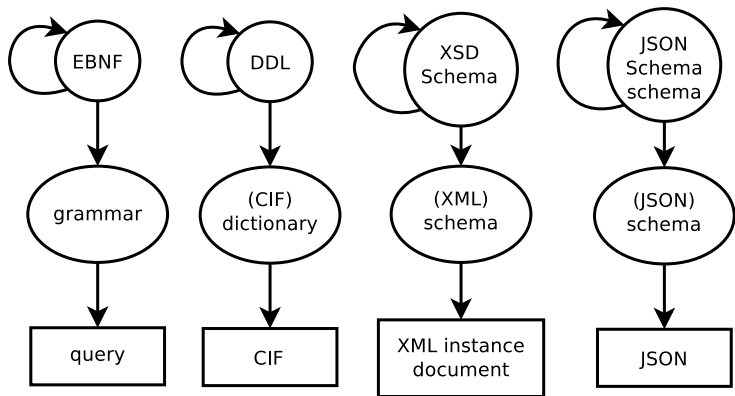
- CIF – dictionaries
- XML, JSON – schemas
- SQL – schemas
- Semantic networks (RDF);

examples/dictionaries/cif-core-example.cif:

```
data_cell_length_
  loop_ _name          '_cell_length_a'
                        '_cell_length_b'
                        '_cell_length_c'
  _category            cell
  _type                numb
  _type_conditions     esd
  _enumeration_range   0.0:
  _units               A
  _units_detail        'angstroms'
  _definition
;      Unit-cell lengths in angstroms corresponding to the structure
      reported. The values of _refln_index_h, *_k, *_l must
      correspond to the cell defined by these values and _cell_angle_
      values. The values of _diffrn_refln_index_h, *_k, *_l may not
      correspond to these values if a cell transformation took place
      following the measurement of the diffraction intensities. See
      also _diffrn_reflns_transf_matrix_.
;
```

Self-describing data

Three levels are always enough!



- IUCr quality criteria
 - IUCr list of data validation criteria
<ftp://ftp.iucr.org/pub/dvntests>
 - IUCr requirements for publications
- COD quality criteria
 - ✓✓ Correct syntax;
 - ✓ Validation using data dictionaries;
 - ✓ Validation using data statistics;
 - ✓ Validation using first physical principles;

IUCr criteria example (<ftp://ftp.iucr.org/pub/dvntests>):

data validation criteria

AUTO CHECK LIST-Version: 2000.06.09

Full list of validation algorithms

CHECK: [CHEMS] _chemical_formula_sum

TESTS:

1. in the order C, H followed by alphabetic [CHEMS_01]
2. classify as organic, inorganic, metal-organic [CHEMS_02]

EXAMPLE(S): 'C18 H19 N7 O8 S'

CHECK: [CHEMW] _chemical_formula_weight

FORMAT: F.2

TESTS:

1. agrees with _chemical_formula_sum [CHEMW_01]
2. agrees with _atom_site_data [CHEMW_03]

COD data validation policy:

① Syntax checks:

```
$ cifparse 7234818.cif
```

② Semantic (dictionary) checks:

```
$ cif_validate -D cif_core.dic 7234818.cif
```

③ Database specific tests:

```
$ cif_cod_check 7234818.cif
```

COD data deposition Web site

Crystallography Open Databas...

Data block 739121:

- » `_journal_name_full` is undefined
- » neither `_journal_year` nor `_journal_volume` is defined
- » `_journal_page_first` is undefined

Tip: if you need to add bibliography common to all structures in this file, you can add a **data_global** section below, and the data will be distributed into all other sections.

Fetch bibliography by DOI (<http://www.doi.org>):

Save and check Fetch Pubmed crossref

Your CIF File contents:

```
data_global
loop
  _publ_author_name
  'Sabiah, Shahulhameed'
  'Lee, Chen-Shiang'
  'Hwang, Wen-Shu'
  'Lin, Ivan J. B.'
  _publ_section_title
;
Facile C-N Bond Cleavage Promoted by Cuprous Oxide: Formation
of C-C-Coupled Bimidazole from Its Methylene-Bridged Congener
;
_journal_issue          2
_journal_name_full      Organometallics
_journal_page_first     290
_journal_volume         29
_journal_year           2010
data_714906
_chemical_formula_sum   'C16 H20 Cl4 Cu2 N8'
_chemical_formula_weight 593.28
```

COD data deposition Web site

Crystallography Open Database: CIF Validator - Mozilla Firefox

http://www.crystallography.net/store.php?f=06&CODSESSION=ZY0lg8DU9KTyEi-KIIS,grO5404

Crystallography Open Database Validation and Deposition Interface

Log in Upload a file **Validate data** Deposit structures Finish

Deposit to COD all valid files

File	Status	Actions
om9010406_si_002.cif	valid	Edit Deposit to COD

File [om9010406_si_002.cif] is correct

- 1 A centralised registry:
 - COD identifiers (e.g. COD 2000000);
 - PDB identifiers (e.g. PDB 1KNV);
 - DOI (e.g. 10.1093/nar/gkn883);
 - ARK (Archival Resource Key, e.g. [ark:/53355/cl010066723](https://n2t.org/ark:/53355/cl010066723));
 - URI (e.g. <https://www.w3.org/Provider/Style/URI.html>);
 - ISSN, ISBN, PMID, PMCID, ...
- 2 Randomised identifiers
 - UUID (e.g. 90376010-a315-11ea-adba-6bb1c61159af)
 - Cryptographic checksums (e.g. git commit 42a03a255612b8d43ecd77bb0acc02def888f688, 42a03a2);

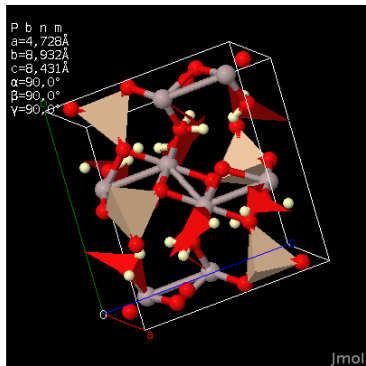
Take home messages

- ① Data provenance is important for reproducible research;
- ② Unique stable identifiers help tracking data;
- ③ Comprehensive metadata are needed for reproducibility;
- ④ Let's use standard well-defined data formats;
- ⑤ Document your computation work flow.

Thank you!



<http://en.wikipedia.org/wiki/Topaz>



Coordinates

[2207377.cif](#)

Original IUCr paper

[HTML](#)

<http://www.crystallography.net/2207377.html>

References I



Fuentes-Cobas, Luis E. et al. (Aug. 2017). “The representation of coupling interactions in the Material Properties Open Database (MPOD)”. In: *Advances in Applied Ceramics* 116.8, pp. 428–433. DOI: 10.1080/17436753.2017.1343782.



Gražulis, Saulius et al. (2009). “Crystallography Open Database – an open-access collection of crystal structures”. In: *Journal of Applied Crystallography* 42, pp. 726–729. DOI: 10.1107/S0021889809016690. URL: <http://dx.doi.org/10.1107/S0021889809016690>.



Gražulis, Saulius et al. (2012). “Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration”. In: *Nucleic Acids Research* 40, pp. D420–D427. DOI: 10.1093/nar/gkr900. URL: <http://nar.oxfordjournals.org/content/40/D1/D420.abstract>.



Hall, S. R. et al. (1991). “The crystallographic information file (CIF): a new standard archive file for crystallography”. In: *Acta Crystallographica Section A* 47, pp. 655–685. DOI: 10.1107/S010876739101067X. URL: <http://dx.doi.org/10.1107/S010876739101067X>.



Li, Chuan-Yun et al. (Jan. 2008). “Genes and (common) pathways underlying drug addiction”. In: *PLoS Computational Biology* 4.1. Ed. by Peter D. Karp, e2. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.0040002. URL: <http://dx.doi.org/10.1371/journal.pcbi.0040002>.

References II



Mendili, Yassine El et al. (May 2019). “Raman Open Database: first interconnected Raman–X-ray diffraction open-access resource for material identification”. In: *Journal of Applied Crystallography* 52.3, pp. 618–625. doi: 10.1107/s1600576719004229.



Pepponi, Giancarlo et al. (2012). “MPOD: A Material Property Open Database linked to structural information”. In: *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms* 284.0. E-MRS 2011 Spring Meeting, Symposium M: X-ray techniques for materials research-from laboratory sources to free electron lasers, pp. 10–14. issn: 0168-583X. doi: 10.1016/j.nimb.2011.08.070. url: <http://www.sciencedirect.com/science/article/pii/S0168583X11008639>.



Senior, Andrew W. et al. (Jan. 2020). “Improved protein structure prediction using potentials from deep learning”. In: *Nature* 577.7792, pp. 706–710. doi: 10.1038/s41586-019-1923-7.



Zheng, Yu et al. (Nov. 2008). “Using shotgun sequence data to find active restriction enzyme genes”. In: *Nucleic Acids Research* 37.1, e1–e1. doi: 10.1093/nar/gkn883. url: <https://doi.org/10.1093/nar/gkn883>.