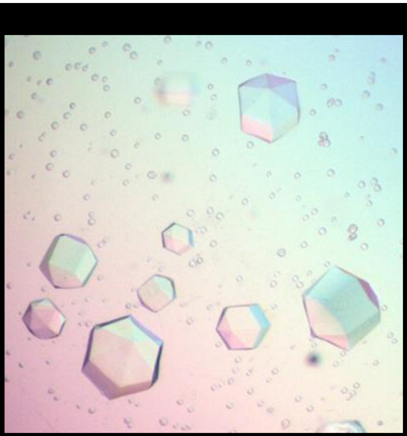
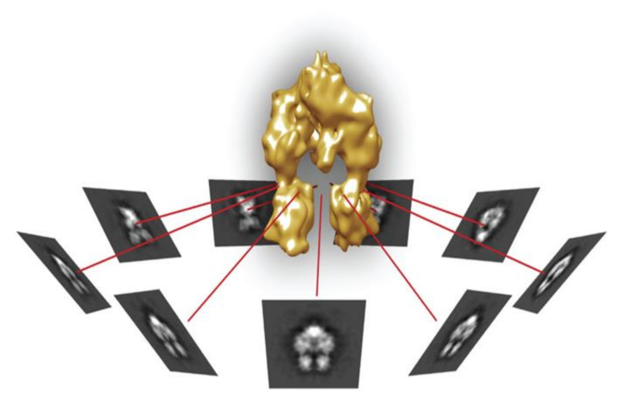
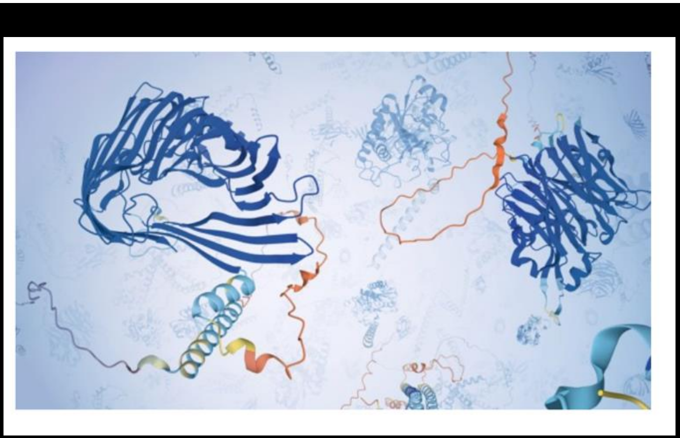






Structural Data:
X-ray Crystallography & Cryo-EM & AI

Jesse Rinehart, PhD

Biomedical Data Science: Mining & Modeling
CBB 752, Spring 2025

Cellular & Molecular Physiology
Yale University School of Medicine

YALE SYSTEMS BIOLOGY INSTITUTE

Yale Structure Courses:

MB&B529b / PHAR529b, Structural Biology and Drug Discovery

MB&B711b / C&MP711b, Practical cryo-EM Workshop

MB&B720a, Macromolecular Structure and Biophysical Analysis

C&MP 710b/MB&B 710b4, Electron Cryo-Microscopy for Protein Structure Determination

MB&B635a / ENAS518a, Quantitative Approaches in Biophysics and Biochemistry

Additional Resources:

"Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models"
Gale Rhodes (Third Edition, 2006 Elsevier/Academic Press)

"Crystallography 101" http://www.ruppweb.org/Xray/101index.html

"Single particle electron cryomicroscopy: trends, issues and future perspective."
Vinothkumar KR, Henderson R. Q Rev Biophys. 2016 pubmed:27658821

"Cryo-EM: A Unique Tool for the Visualization of Macromolecular Complexity"
Eva Nogales & Sjors HW Scheres, Mol. Cell 015 May PMID: 26000851

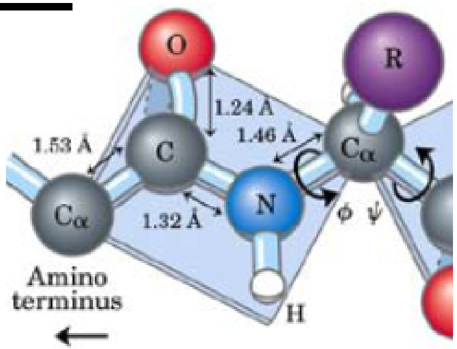
Thank you to **Yong Xiong** and **Fred Sigworth** for contributions to this lecture

“Just as we see objects around us by interpreting the light reflected from them, x-ray crystallographers "see" molecules by interpreting x-rays diffracted from them.”
- Gale Rhodes

- There's a <u>limit</u> to how small an object can be seen under a light microscope.
- <u>The diffraction limit</u>: you can not image things that are much smaller than the wavelength of the light you are using.
- The wavelength for visible light is measured in hundreds of nanometers, while atoms are separated by distances of the order of 0.1nm, or 1Å.

<u>We need to use x-rays to resolve atomic features.</u>

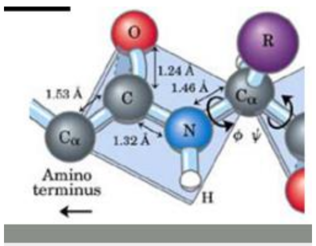
Distances between atoms are small:
Lab x-ray sources use CuKα radiation. Wavelength = 1.54 Å.
Synchrotron radiation wavelengths in the range 0.5 Å - 2.5 Å.



Yong Xiong

The 2014 Nobel Prize in Chemistry: Eric Betzig, W.E. Moerner, and Stefan Hell
“The development of super-resolved fluorescence microscopy"

1Å = 0.1nm



Spatial Resolution of Biological Imaging Techniques

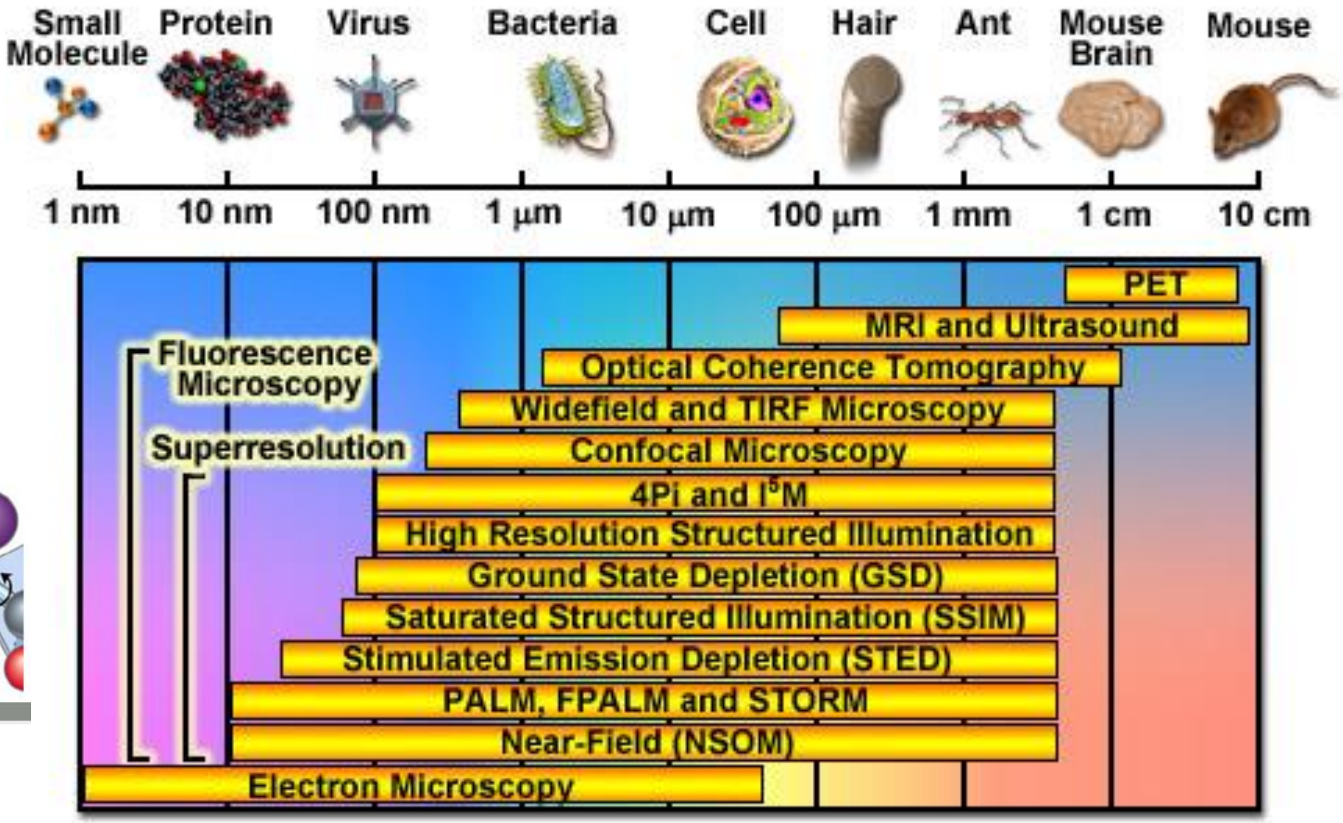
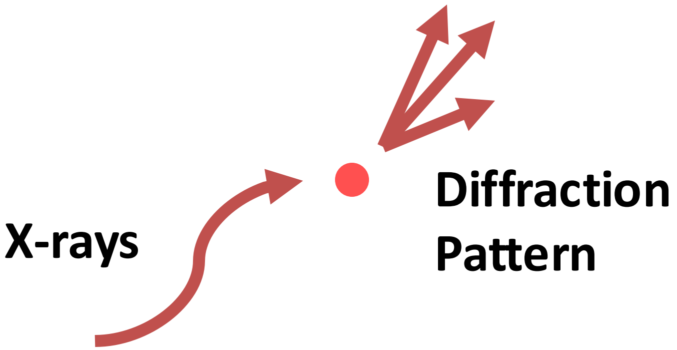


Figure 1

http://zeiss-campus.magnet.fsu.edu/articles/superresolution/introduction.html

Experimental Determination of Atomic Resolution Structures

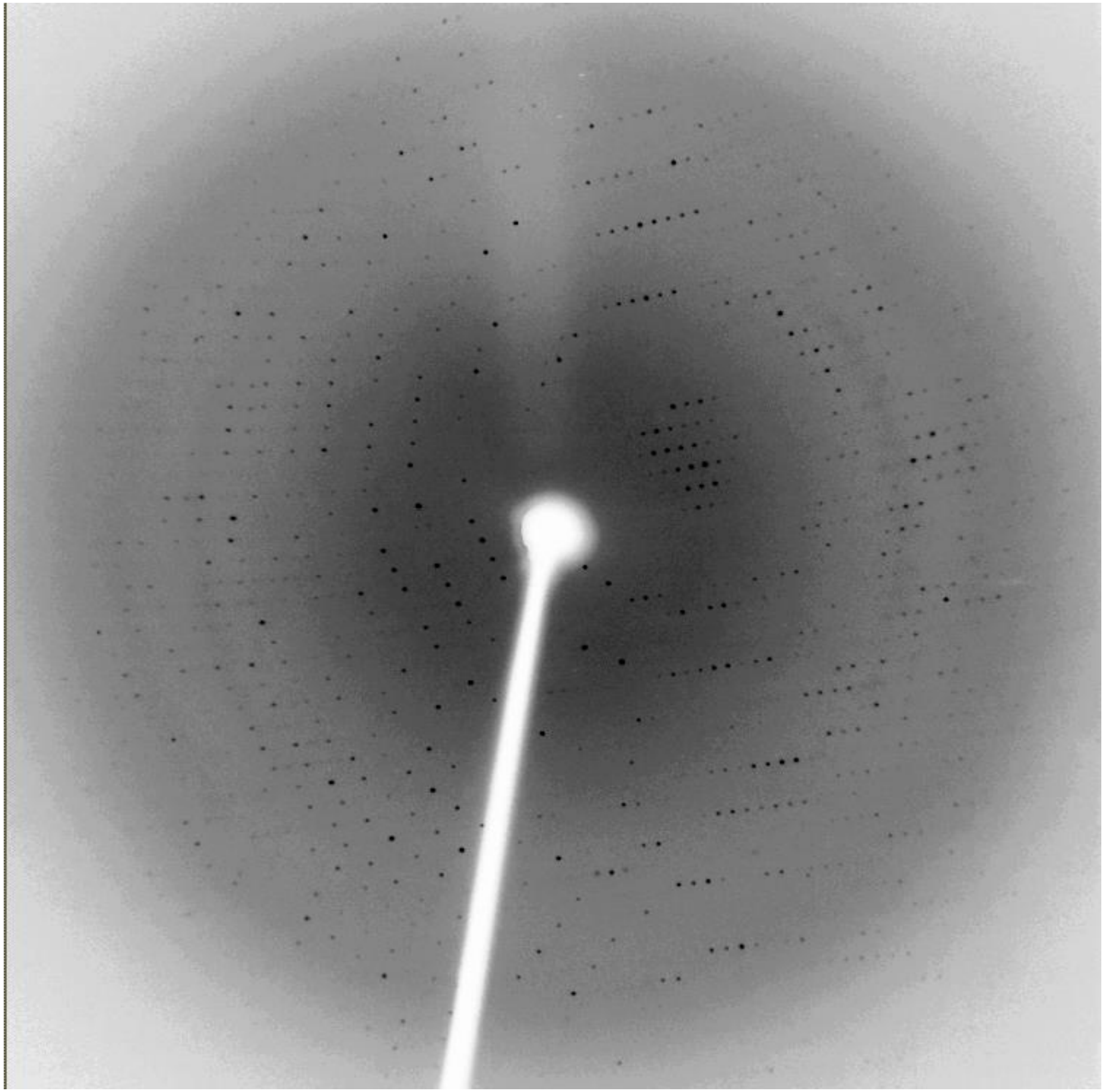
X-ray



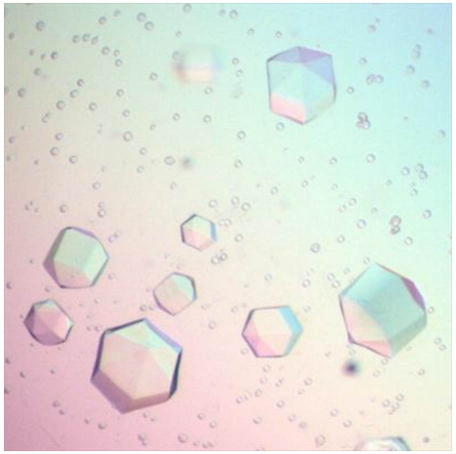
- **Direct detection of atom positions**
- **Crystals required**

Other methods for determining protein structures:
-EM (Electron Microscopy), Cryo-EM, ESR/Fluorescence

Image of X-ray diffraction of a protein crystal



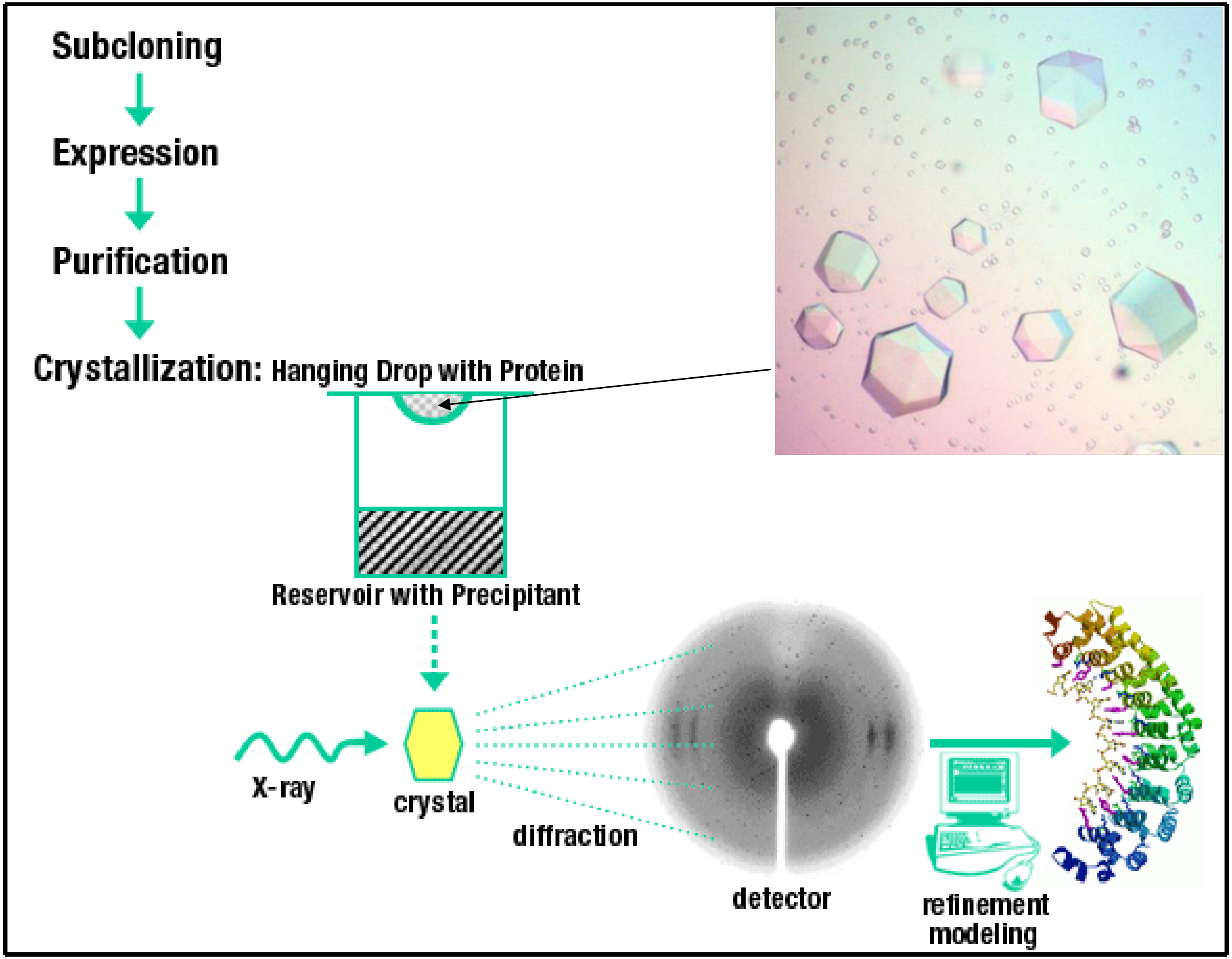
Why Crystals?



X-rays are scattered by electrons, too weak to record scattering from a single molecule. Crystals are therefore used because they present many molecules (N) in exactly the same orientation. The scattering from each of the N molecules interferes constructively to give a measurable diffraction pattern (enhanced ~N^2 fold).

Yong Xiong

Determination of Protein Crystal Structure



Subcloning

Expression

Purification

Crystallization: Hanging Drop with Protein

Reservoir with Precipitant

X-ray

crystal

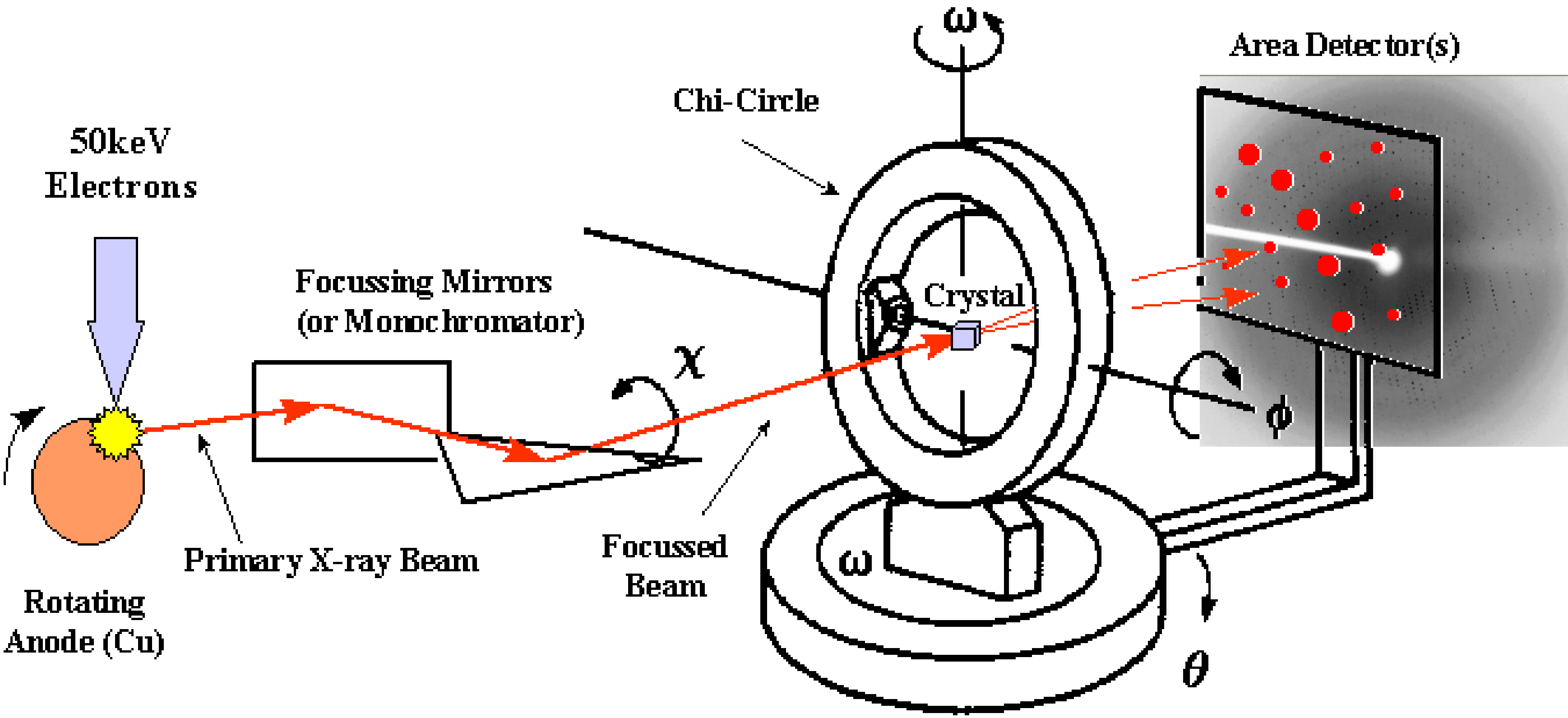
diffraction

detector

refinement
modeling

http://www.noble.org/PlantBio/Wang/crystallography.html

Data Collection

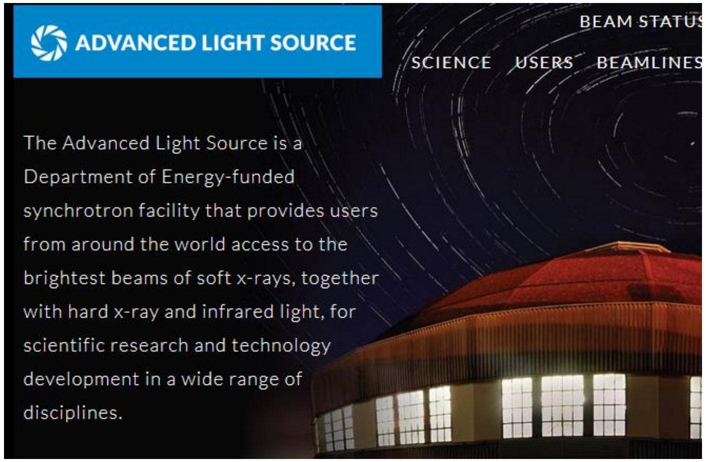


4-Circle Gonoimeter (Eulerian or Kappa Geometry)

Synchrotron X-ray Sources are the method of choice

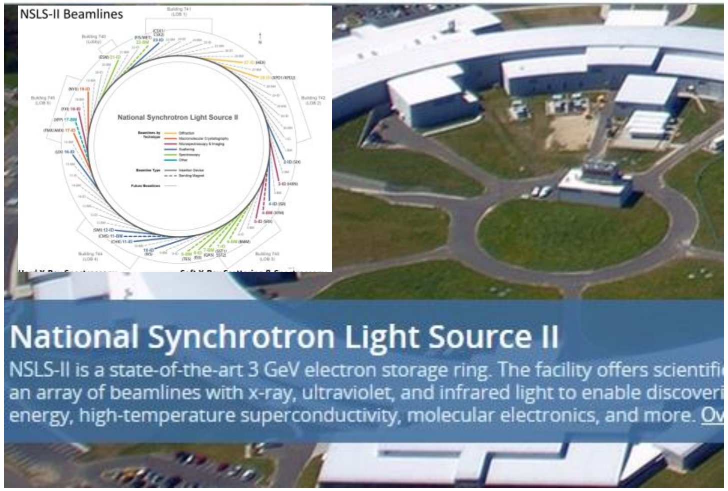
Lab x-ray sources @ 1.54 Å compared to Synchrotron X-ray @ 0.5 Å - 2.5 Å.

APS Chicago



The Advanced Light Source is a Department of Energy-funded synchrotron facility that provides users from around the world access to the brightest beams of soft x-rays, together with hard x-ray and infrared light, for scientific research and technology development in a wide range of disciplines.

ALS Berkeley



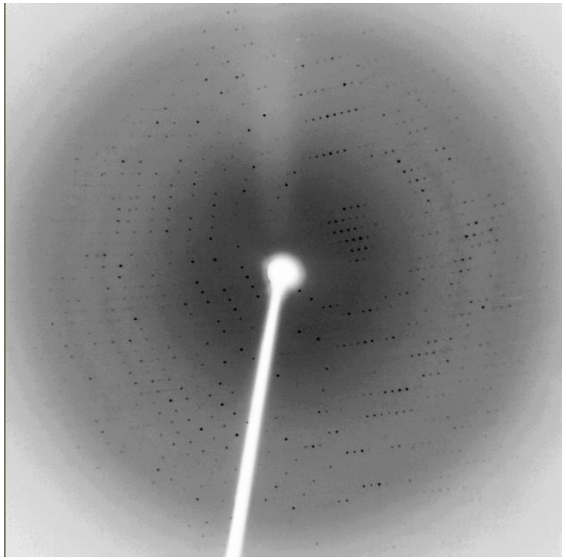
National Synchrotron Light Source II

NSLS-II is a state-of-the-art 3 GeV electron storage ring. The facility offers scientifi an array of beamlines with x-ray, ultraviolet, and infrared light to enable discoveri energy, high-temperature superconductivity, molecular electronics, and more. Ov

NSLS-II Brookhaven

CHESS Ithaca

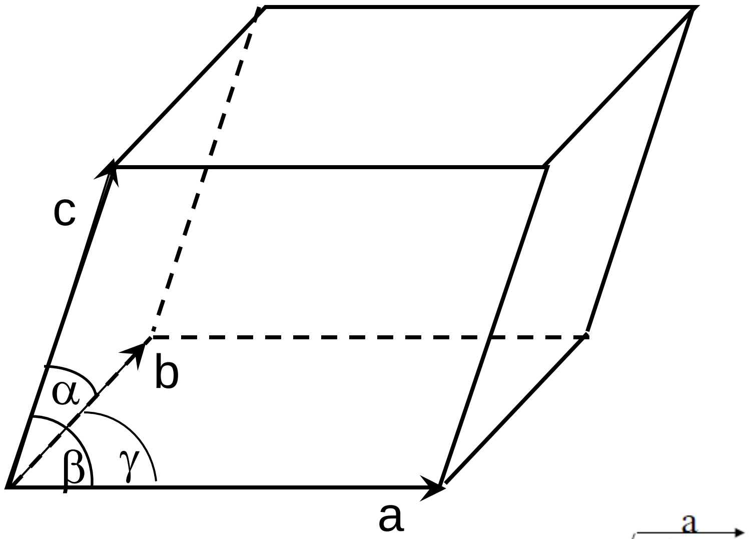
The information we get from a single diffraction experiment

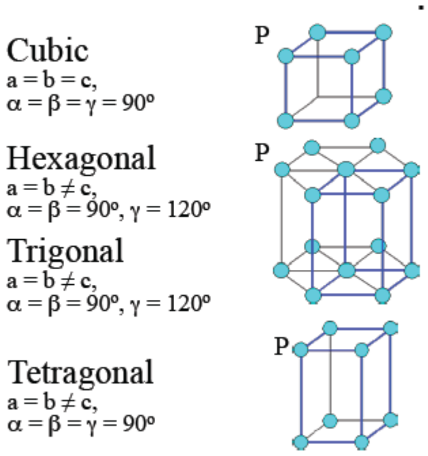


Analyze the pattern of the reflections →

(a) space group of the crystal

(b) unit cell dimensions





Cubic
a = b = c,
α = β = γ = 90º

Hexagonal
a = b ≠ c,
α = β = 90º, γ = 120º

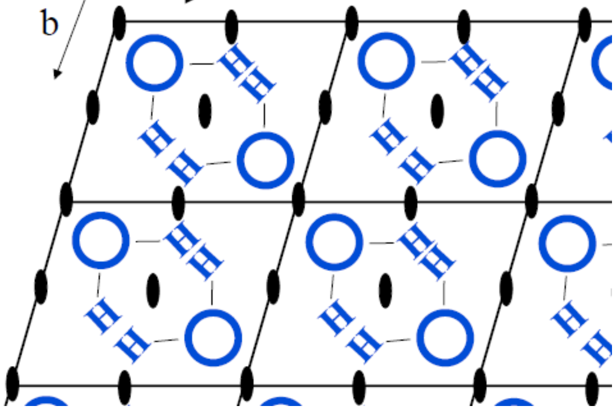
Trigonal
a = b ≠ c,
α = β = 90º, γ = 120º

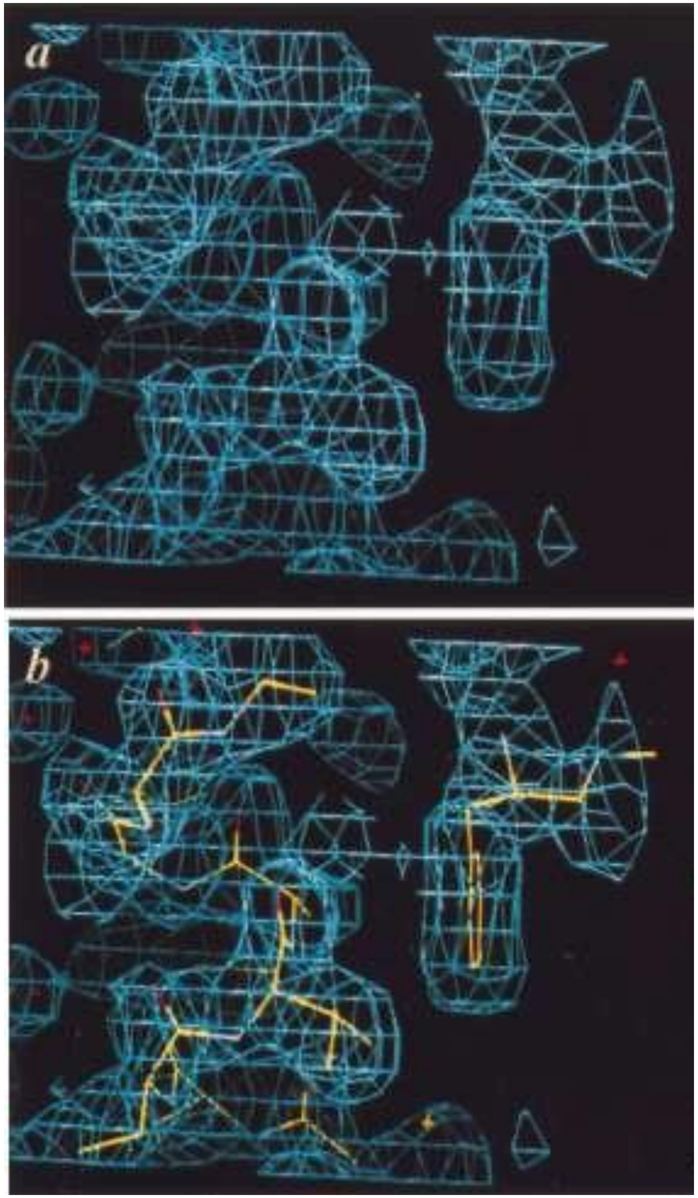
Tetragonal
a = b ≠ c,
α = β = γ = 90º

How to understand symmetry?

Crystal = lattice + unit cell content

(asymmetric units (asu) content)

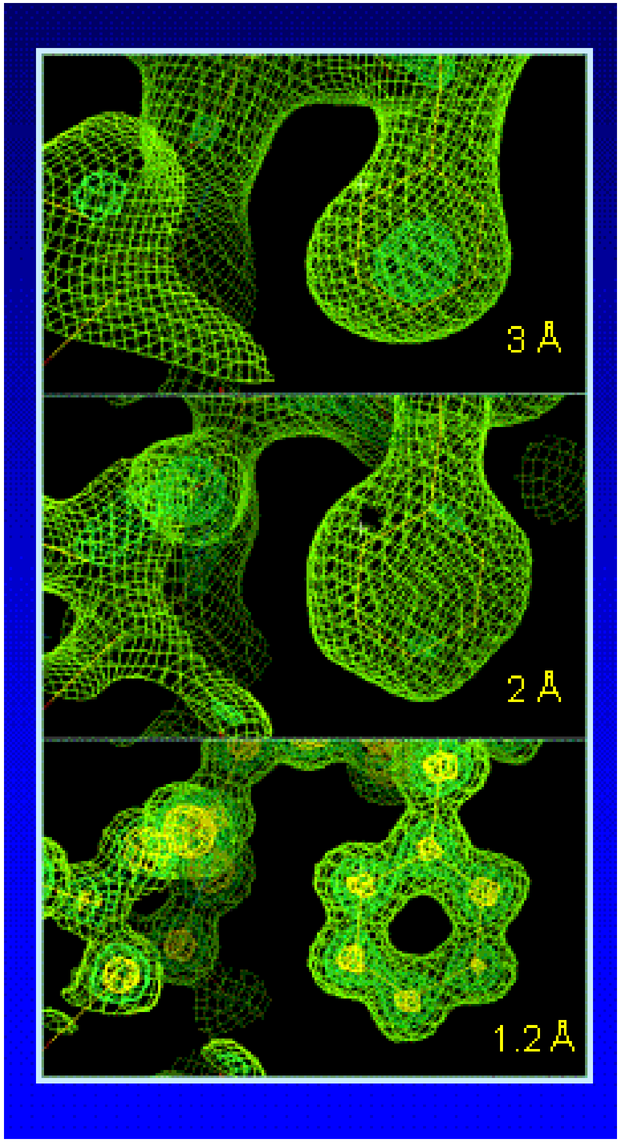


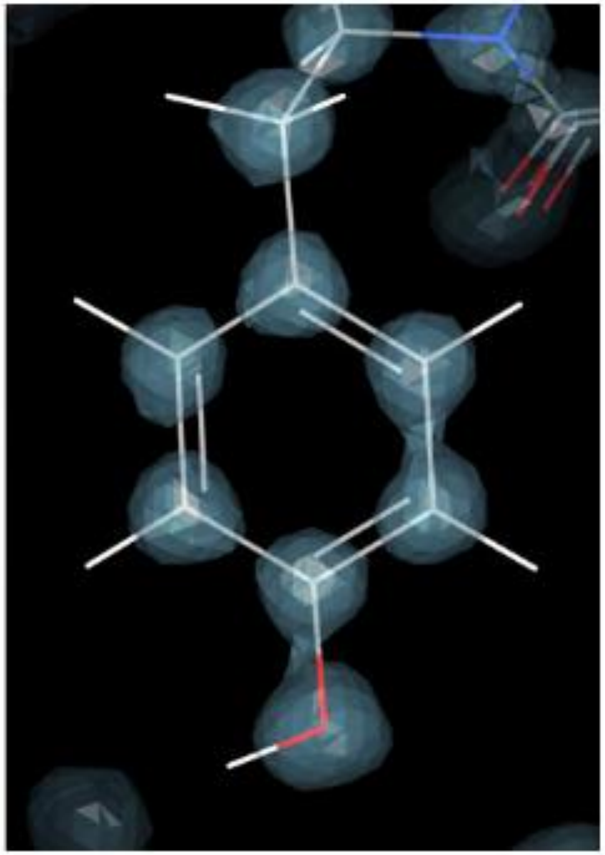


Electron density map

Building a structure model

The importance of resolution

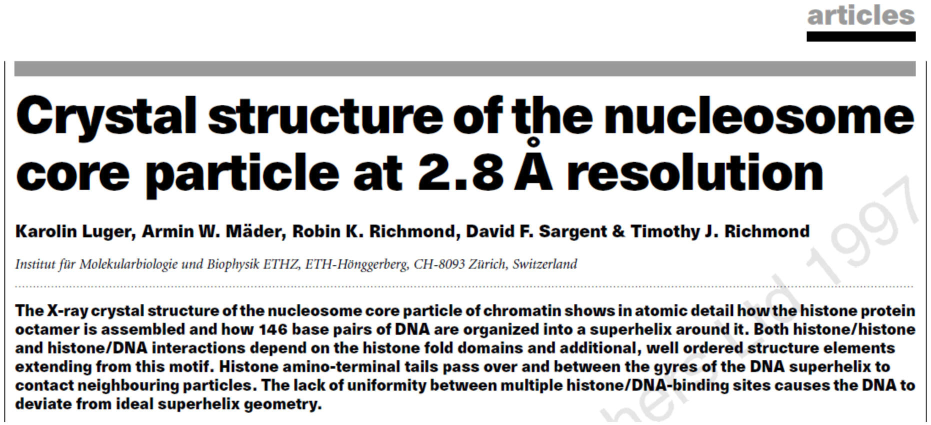




Crystal structure of small protein crambin at 0.48 A resolution
Schmidt, A., et al (2011) Acta Crystallography 67: 424-429

https://www.rcsb.org/structure/3nir

http://www.ruppweb.org/Xray/101index.html

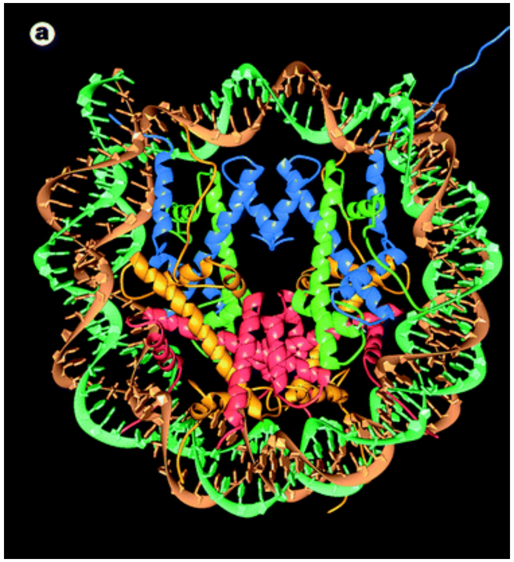


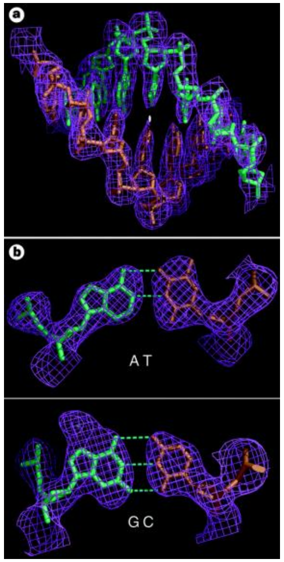
Crystal structure of the nucleosome core particle at 2.8 Å resolution

Karolin Luger, Armin W. Mäder, Robin K. Richmond, David F. Sargent & Timothy J. Richmond

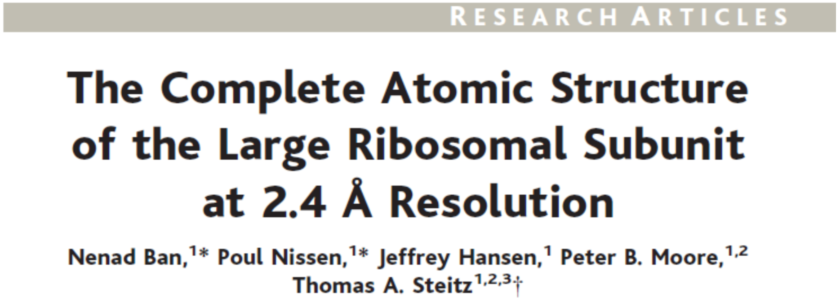
Institut für Molekularbiologie und Biophysik ETHZ, ETH-Hönggerberg, CH-8093 Zürich, Switzerland

The X-ray crystal structure of the nucleosome core particle of chromatin shows in atomic detail how the histone protein octamer is assembled and how 146 base pairs of DNA are organized into a superhelix around it. Both histone/histone and histone/DNA interactions depend on the histone fold domains and additional, well ordered structure elements extending from this motif. Histone amino-terminal tails pass over and between the gyres of the DNA superhelix to contact neighbouring particles. The lack of uniformity between multiple histone/DNA-binding sites causes the DNA to deviate from ideal superhelix geometry.



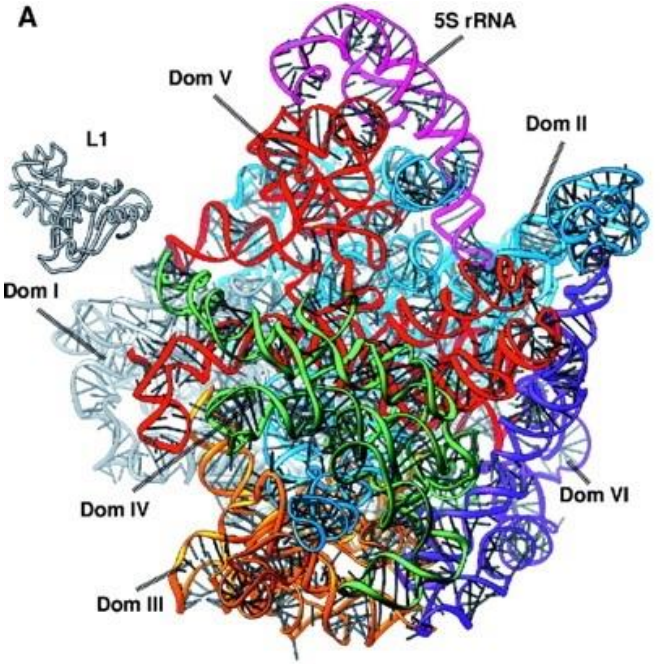


- PMID: 9305837



The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 Å Resolution

Nenad Ban,[1*] Poul Nissen,[1*] Jeffrey Hansen,[1] Peter B. Moore,[1,2] Thomas A. Steitz[1,2,3†]

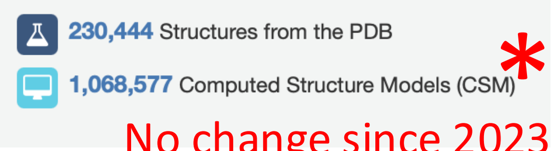


Thomas Steitz shared 2009 Nobel Prize in Chemistry for this structure

Structure Databases

- Where does protein structural information reside?
  - PDB:
    - http://www.rcsb.org/pdb/
  - MMDB:
    - http://www.ncbi.nlm.nih.gov/Structure/
  - FSSP:
    - http://www.ebi.ac.uk/dali/fssp/
  - SCOP:
    - http://scop.mrc-lmb.cam.ac.uk/scop/
  - CATH:
    - http://www.biochem.ucl.ac.uk/bsm/cath_new/

```
# of PDB structures
  2023:    200,988
  2025:    230,444
```

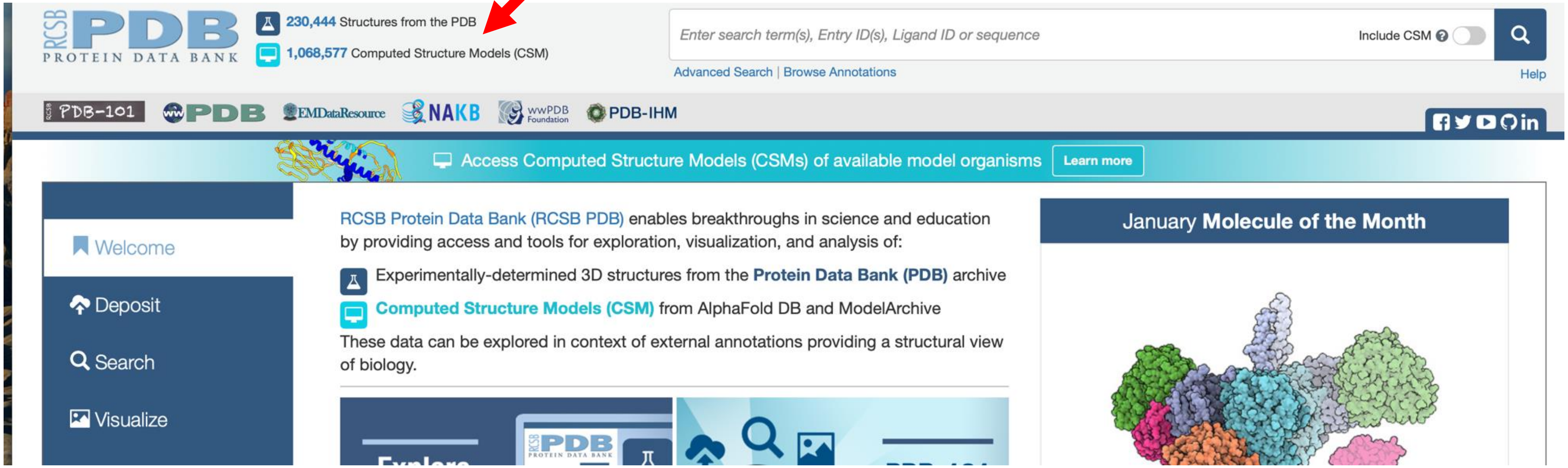


230,444 Structures from the PDB

1,068,577 Computed Structure Models (CSM)

*

No change since 2023



230,444 Structures from the PDB

1,068,577 Computed Structure Models (CSM)

Enter search term(s), Entry ID(s), Ligand ID or sequence

Include CSM

Access Computed Structure Models (CSMs) of available model organisms

RCSB Protein Data Bank (RCSB PDB) enables breakthroughs in science and education by providing access and tools for exploration, visualization, and analysis of:

Experimentally-determined 3D structures from the Protein Data Bank (PDB) archive

Computed Structure Models (CSM) from AlphaFold DB and ModelArchive

These data can be explored in context of external annotations providing a structural view of biology.

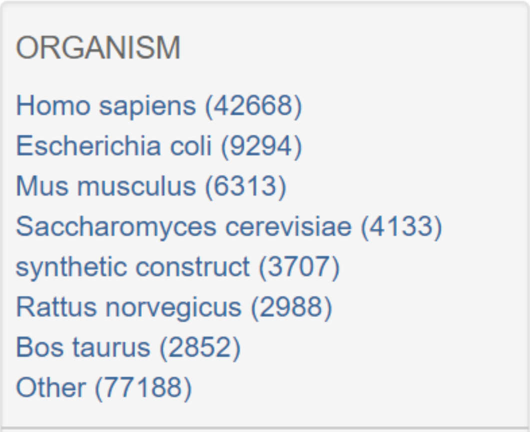
January Molecule of the Month

Updated resources for exploring experimentally-determined PDB structures and Computed Structure Models at the RCSB Protein Data Bank. Nucleic Acids Research, January 2025, Pages D1–D9, https://doi.org/10.1093/nar/gkae1220

https://pdb101.rcsb.org/learn/videos/what-is-a-protein-video

PDB: What species are the structures from?

human

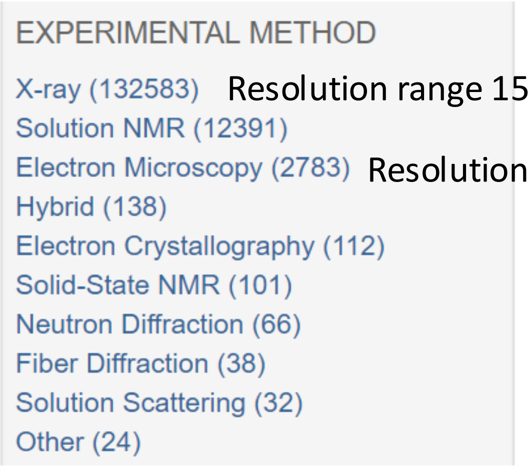


ORGANISM

Homo sapiens (42668)
Escherichia coli (9294)
Mus musculus (6313)
Saccharomyces cerevisiae (4133)
synthetic construct (3707)
Rattus norvegicus (2988)
Bos taurus (2852)
Other (77188)

Which methods?

X-ray



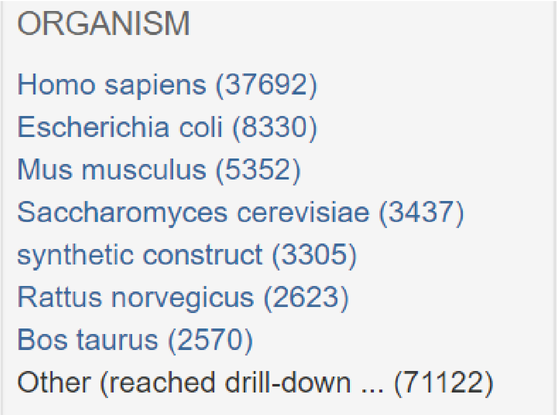
EXPERIMENTAL METHOD

X-ray (132583) Resolution range 15 - 0.48 Å
Solution NMR (12391)
Electron Microscopy (2783) Resolution range 70 - 1.8 Å
Hybrid (138)
Electron Crystallography (112)
Solid-State NMR (101)
Neutron Diffraction (66)
Fiber Diffraction (38)
Solution Scattering (32)
Other (24)

http://www.rcsb.org/pdb/home/home.do

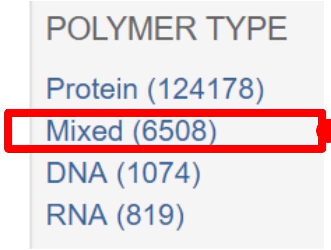
PDB X-ray Structures:

http://www.rcsb.org/pdb/results/results.do?outformat=&qrid=1B04C26E&tabtoshow=Current



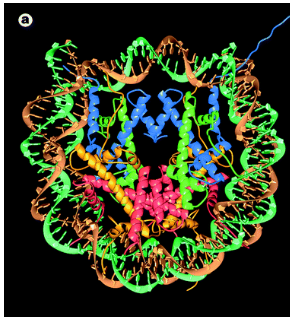
ORGANISM

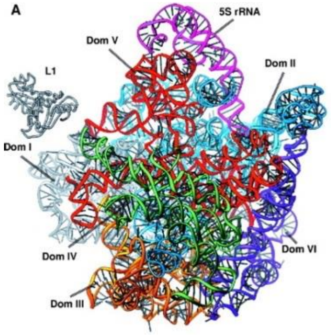
Homo sapiens (37692)
Escherichia coli (8330)
Mus musculus (5352)
Saccharomyces cerevisiae (3437)
synthetic construct (3305)
Rattus norvegicus (2623)
Bos taurus (2570)
Other (reached drill-down ... (71122)

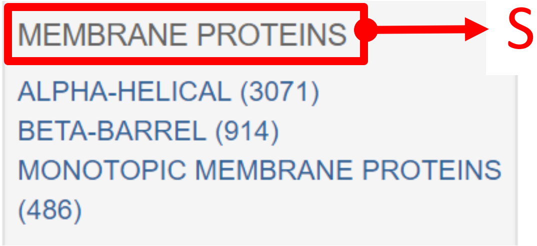


POLYMER TYPE

Protein (124178)
Mixed (6508)
DNA (1074)
RNA (819)







MEMBRANE PROTEINS

ALPHA-HELICAL (3071)
BETA-BARREL (914)
MONOTOPIC MEMBRANE PROTEINS (486)

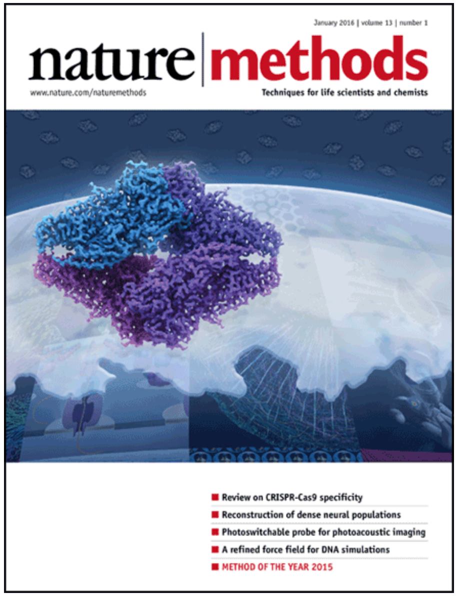
Small % of the total x-ray data

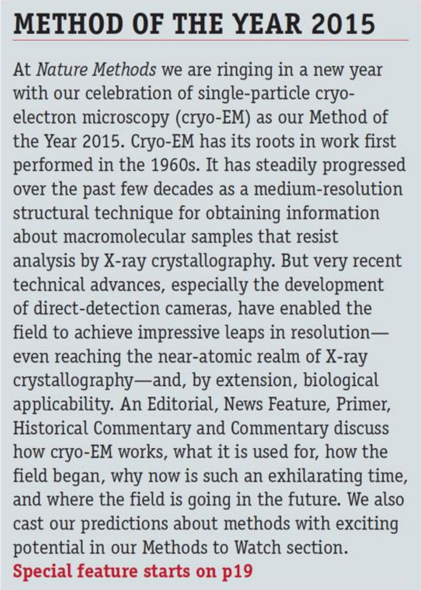
Tools for Viewing Structures

- Jmol
  - http://jmol.sourceforge.net
- PyMOL
  - http://pymol.sourceforge.net
- Swiss PDB viewer
  - http://www.expasy.ch/spdbv
- Mage/KiNG
  - http://kinemage.biochem.duke.edu/software/mage.php
  - http://kinemage.biochem.duke.edu/software/king.php
- Rasmol
  - http://www.umass.edu/microbio/rasmol/

Cryo-EM for biomolecular structures

2015 Method of the Year: Single-particle Cryo-EM


January 2016 | volume 13 | number 1
nature methods
www.nature.com/naturemethods
Techniques for life scientists and chemists
Review on CRISPR-Cas9 specificity
Reconstruction of dense neural populations
Photoswitchable probe for photoacoustic imaging
A refined force field for DNA simulations
METHOD OF THE YEAR 2015



METHOD OF THE YEAR 2015

At *Nature Methods* we are ringing in a new year with our celebration of single-particle cryo-electron microscopy (cryo-EM) as our Method of the Year 2015. Cryo-EM has its roots in work first performed in the 1960s. It has steadily progressed over the past few decades as a medium-resolution structural technique for obtaining information about macromolecular samples that resist analysis by X-ray crystallography. But very recent technical advances, especially the development of direct-detection cameras, have enabled the field to achieve impressive leaps in resolution—even reaching the near-atomic realm of X-ray crystallography—and, by extension, biological applicability. An Editorial, News Feature, Primer, Historical Commentary and Commentary discuss how cryo-EM works, what it is used for, how the field began, why now is such an exhilarating time, and where the field is going in the future. We also cast our predictions about methods with exciting potential in our Methods to Watch section.

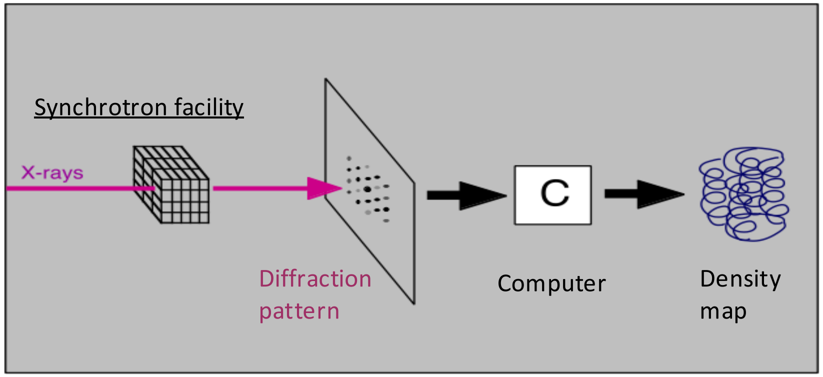
Special feature starts on p19

2017 Nobel Prize in Chemistry

"for developing cryo-electron microscopy for the high-resolution structure determination of biomolecules in solution"

Jacques Dubochet (University of Lausanne, Switzerland)
Joachim Frank (Columbia University, New York, USA)
Richard Henderson (MRC Laboratory of Molecular Biology, Cambridge, UK)

Two methods for structure determination

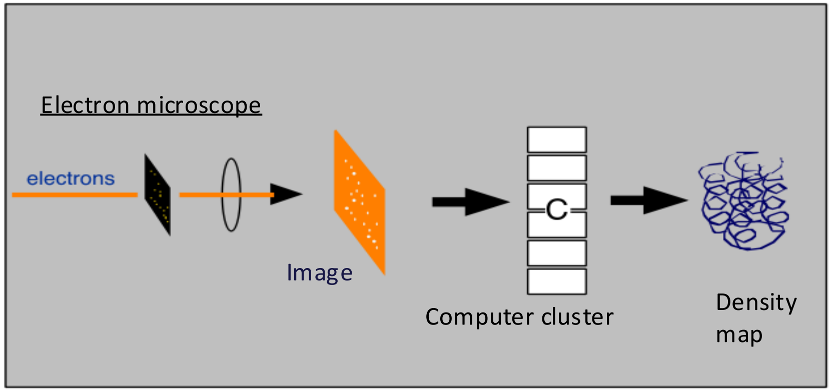


X-ray crystallography

Well-established (since 1960s)

Requires well-ordered crystals

>10^{12} copies of protein



Single-particle cryo-EM

Recent (1990s-present)

No crystals required!

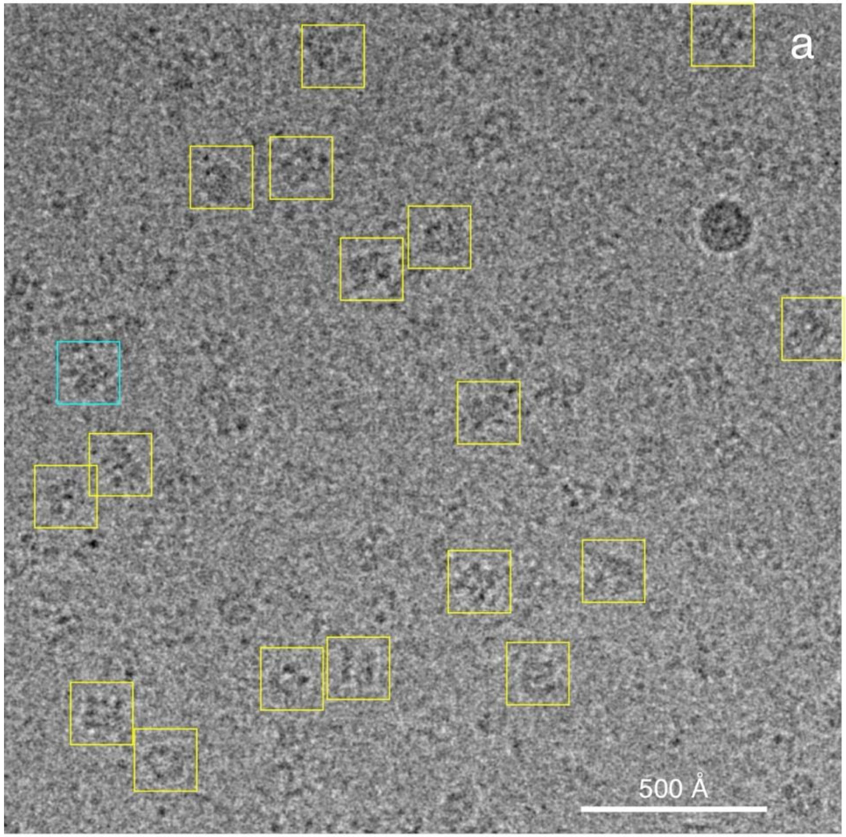
~10^{5} copies of protein

Fred Sigworth

The Cryo-EM specimen gives only a phase contrast image

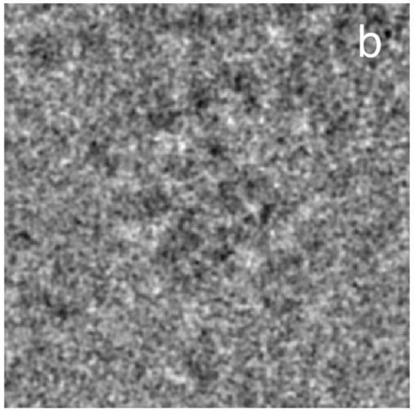
- A constellation of images and data processing are essential.

1/4 of a micrograph, showing some particles

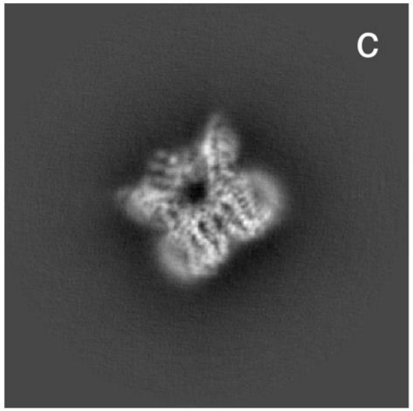


Y. Cheng and D. Julius lab. Nature 2013

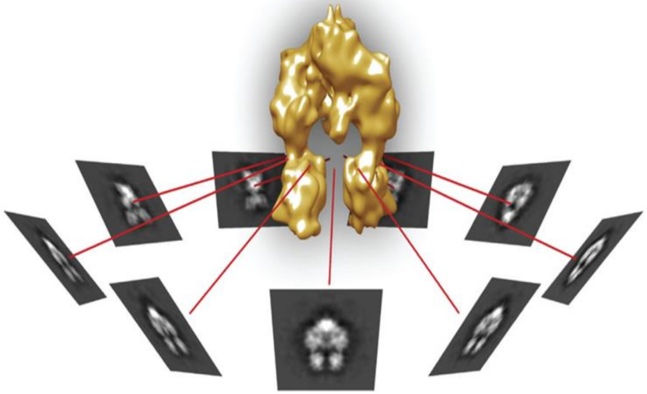
Image



Projection



- orientation assignment and averaging
- 3D reconstruction

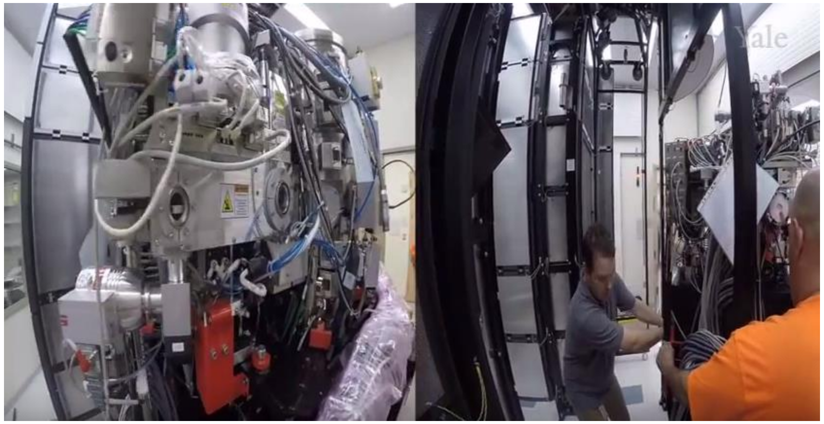


Fred Sigworth

New Technologies, Automation, & Computation are accelerating the field

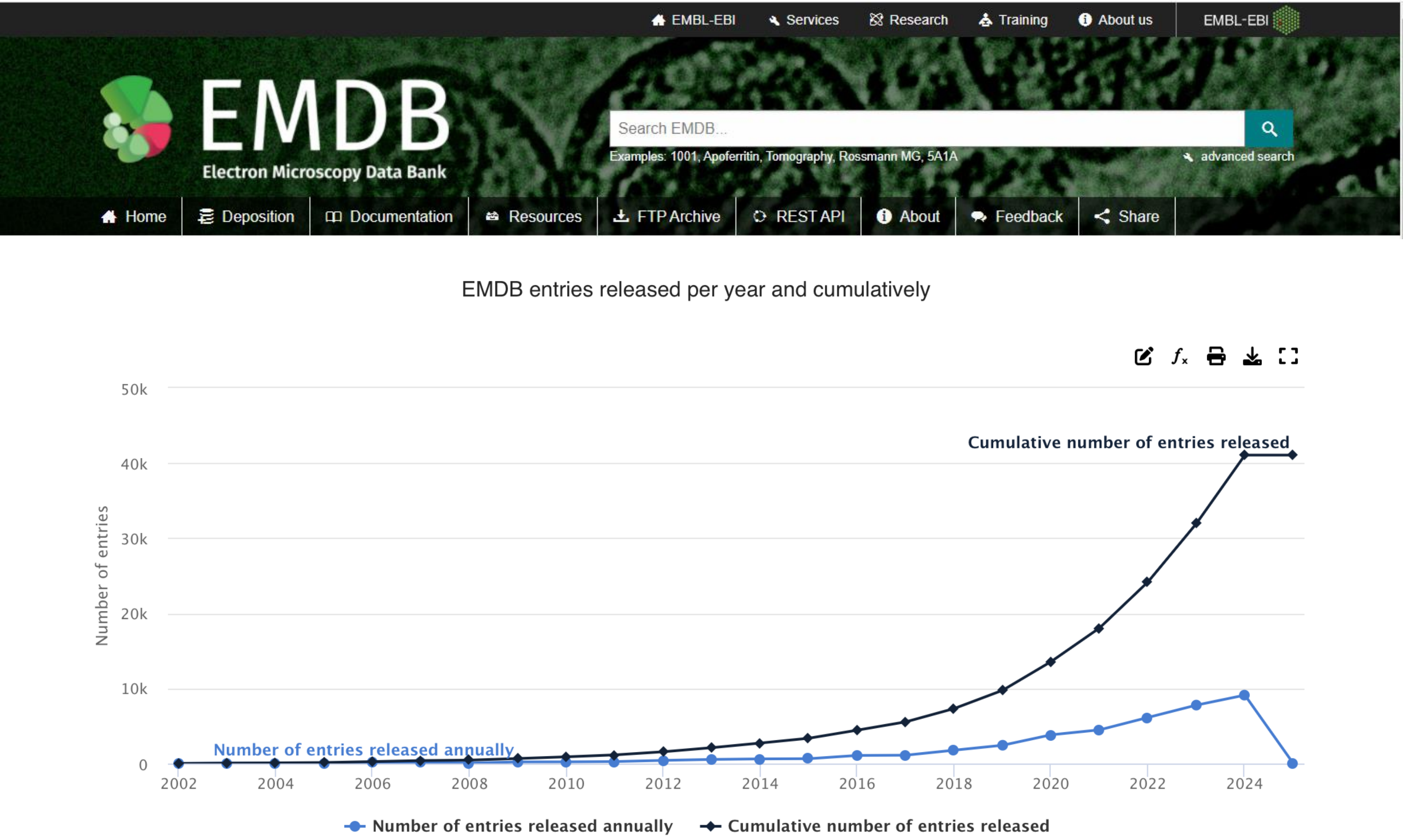
Krios at National University of Singapore

Control room at Scripps Research Institute, La Jolla



Krios TEM installation on Yale's West Campus.

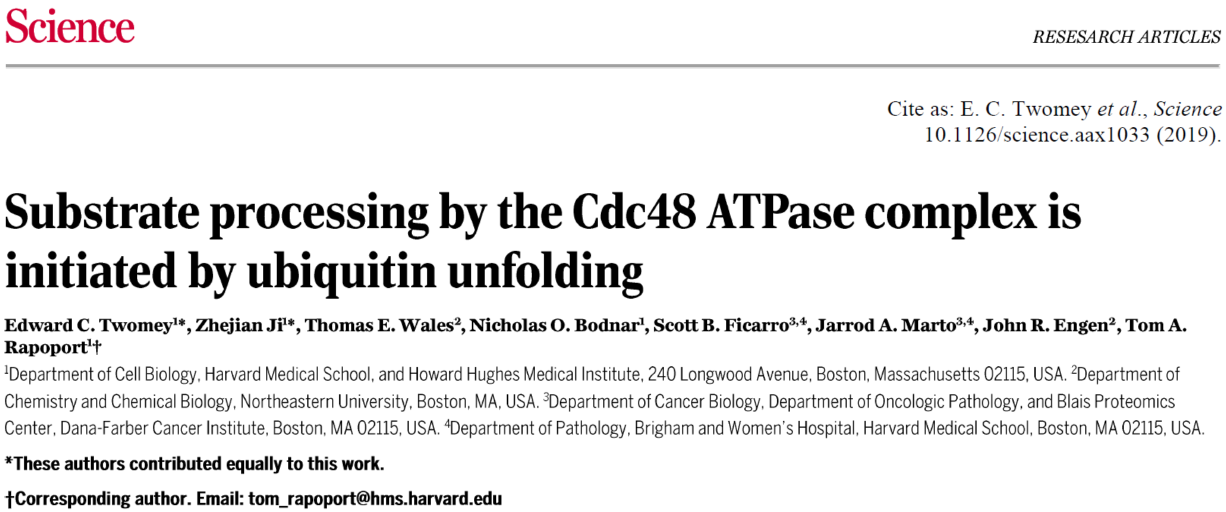
Fred Sigworth



EMDB entries released per year and cumulatively

https://wwwdev.ebi.ac.uk/emdb/statistics

Cryo-EM: membrane proteins, protein complexes, proteins difficult to crystalize



Science

RESEASRCH ARTICLES

Cite as: E. C. Twomey *et al.*, *Science* 10.1126/science.aax1033 (2019).

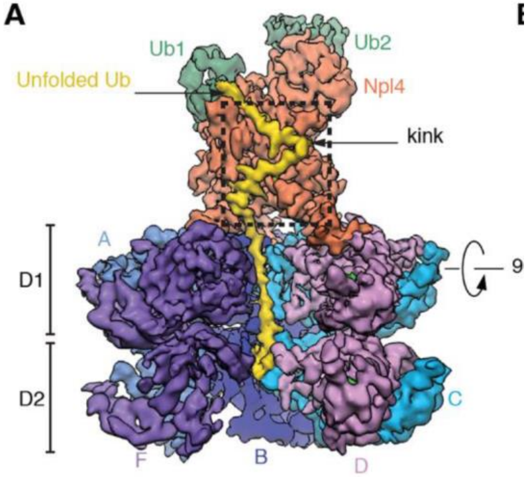
Substrate processing by the Cdc48 ATPase complex is initiated by ubiquitin unfolding

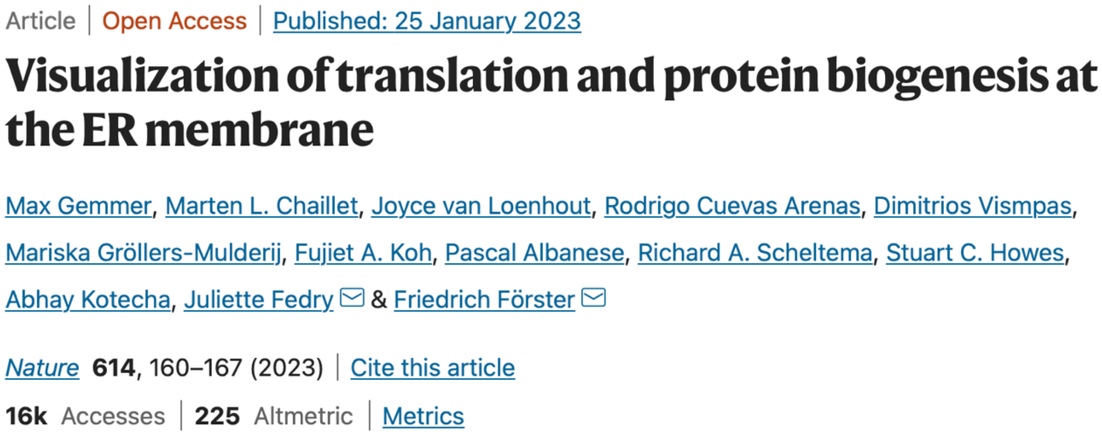
Edward C. Twomey[1]*, Zhejian Ji[1]*, Thomas E. Wales[2], Nicholas O. Bodnar[1], Scott B. Ficarro[3,4], Jarrod A. Marto[3,4], John R. Engen[2], Tom A. Rapoport[1]†

[1]Department of Cell Biology, Harvard Medical School, and Howard Hughes Medical Institute, 240 Longwood Avenue, Boston, Massachusetts 02115, USA. [2]Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA, USA. [3]Department of Cancer Biology, Department of Oncologic Pathology, and Blais Proteomics Center, Dana-Farber Cancer Institute, Boston, MA 02115, USA. [4]Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA.

***These authors contributed equally to this work.**

†Corresponding author. Email: tom_rapoport@hms.harvard.edu





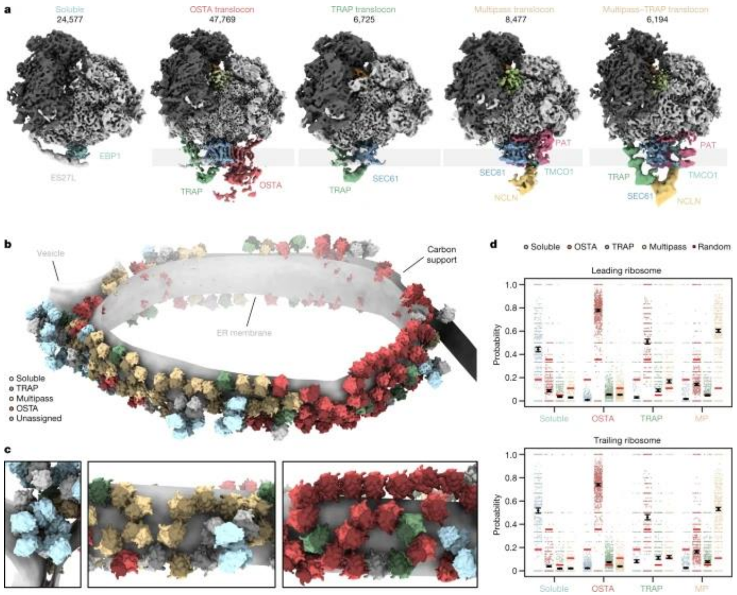
Article | Open Access | Published: 25 January 2023

Visualization of translation and protein biogenesis at the ER membrane

Max Gemmer, Marten L. Chaillet, Joyce van Loenhout, Rodrigo Cuevas Arenas, Dimitrios Vismpas, Mariska Gröllers-Mulderij, Fujiet A. Koh, Pascal Albanese, Richard A. Scheltema, Stuart C. Howes, Abhay Kotecha, Juliette Fedry ✉ & Friedrich Förster ✉

Nature **614**, 160–167 (2023) | Cite this article

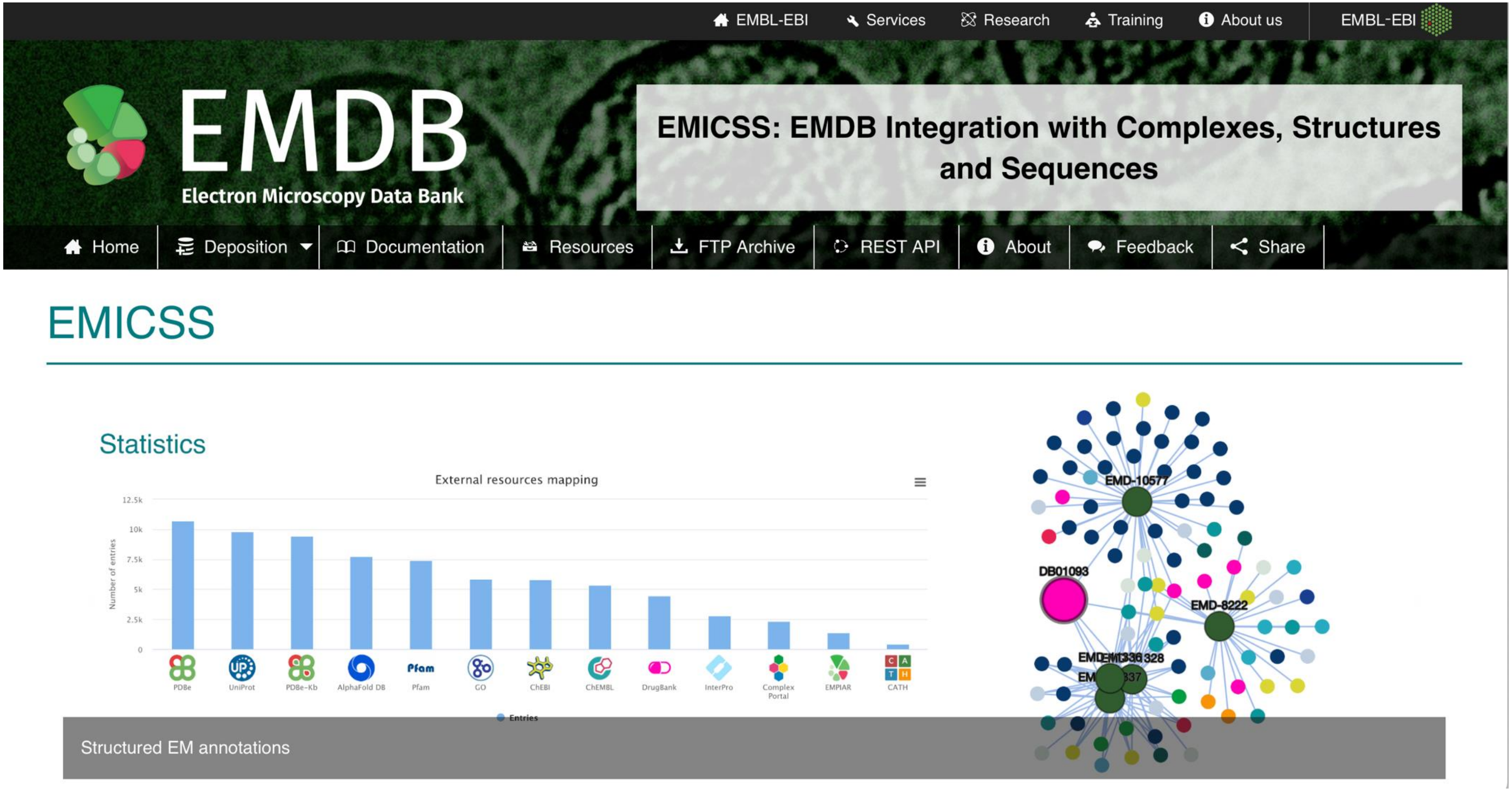
16k Accesses | **225** Altmetric | Metrics



PMID: 30630874;30598546;25918421;31249135;36697828

EMICSS (Launched Dec 2022)

EMDB Integration with Complexes, Structures and Sequences.



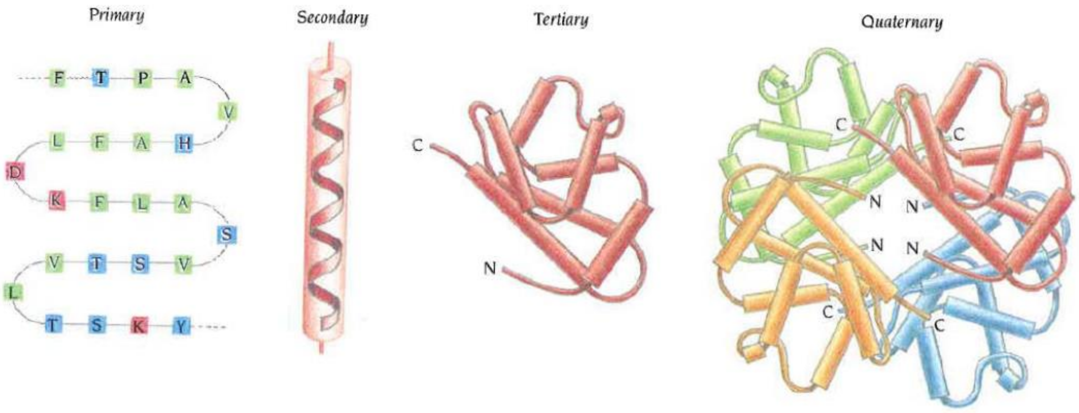
This service provides weekly updated cross-reference information for all EMDB entries, including both entry-level annotations (e.g., publication, corresponding PDB and EMPIAR entries, etc.) and sample-level (e.g., UniProt identifiers, AlphaFold DB models, etc.) annotations. The information from EMICSS is used on the EMDB website to provide relevant links and annotation for individual entries and sample components. The search system also takes advantage of this data to enable advanced queries not otherwise possible.

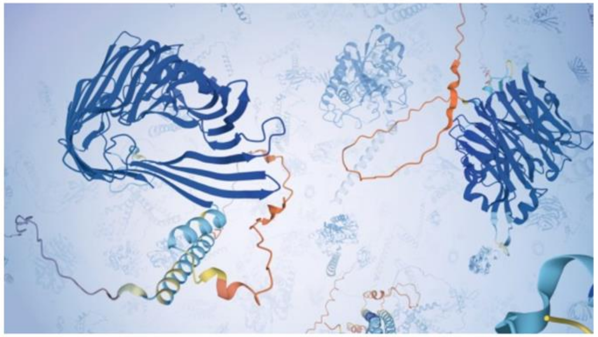
https://www.ebi.ac.uk/emdb/emicss

The protein-folding problem was first posed over 50 years ago:

What is the physical code by which an amino acid sequence dictates fold?

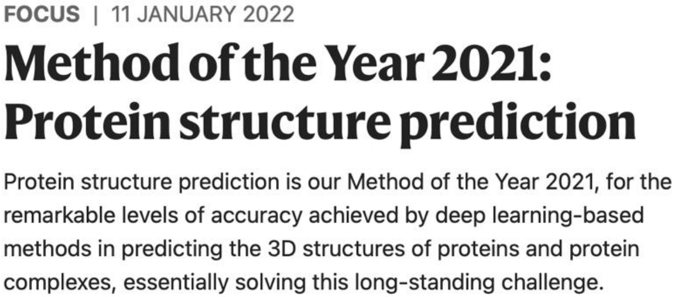
Can we devise a computer algorithm to predict protein structures from their sequences?





The Protein-Folding Problem, 50 Years On, Dill K and Maccallum, J.L. Science, 2012, PMID: 23180855
Proteins and Protein Structure (Branden, C. and Tooze, J. *Introduction to Protein Structure*)

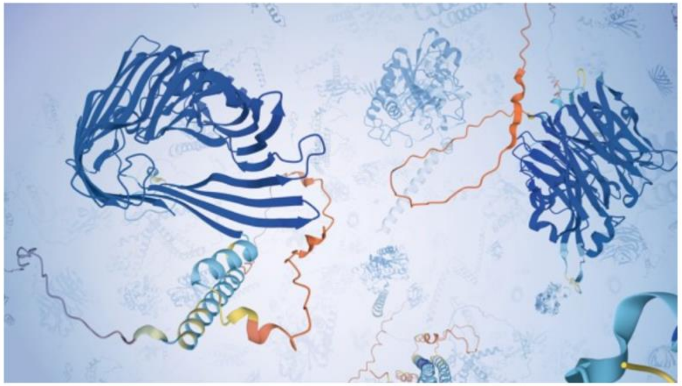
AI deep-learning-based methods solved the protein folding problem



FOCUS | 11 JANUARY 2022

Method of the Year 2021: Protein structure prediction

Protein structure prediction is our Method of the Year 2021, for the remarkable levels of accuracy achieved by deep learning-based methods in predicting the 3D structures of proteins and protein complexes, essentially solving this long-standing challenge.



Predicting and designing protein structures wins a 2024 Nobel Prize

David Baker (left) figured out how to build new proteins. Demis Hassabis (middle) and John Jumper (right) developed an AI tool to predict protein structures.
NIKLAS ELMEHED, © NOBEL PRIZE OUTREACH

Key literature:

(AlphaFold)

Senior, A. W. et al. *Nature* **577**, 706–710 (2020). PMID: 34293799.

Jumper, J. et al. *Nature* **596**, 583–589 (2021). PMID: 34265844.

Tunyasuvunakool, K. et al. *Nature* **596**, 590–596 (2021) PMID: 34293799.

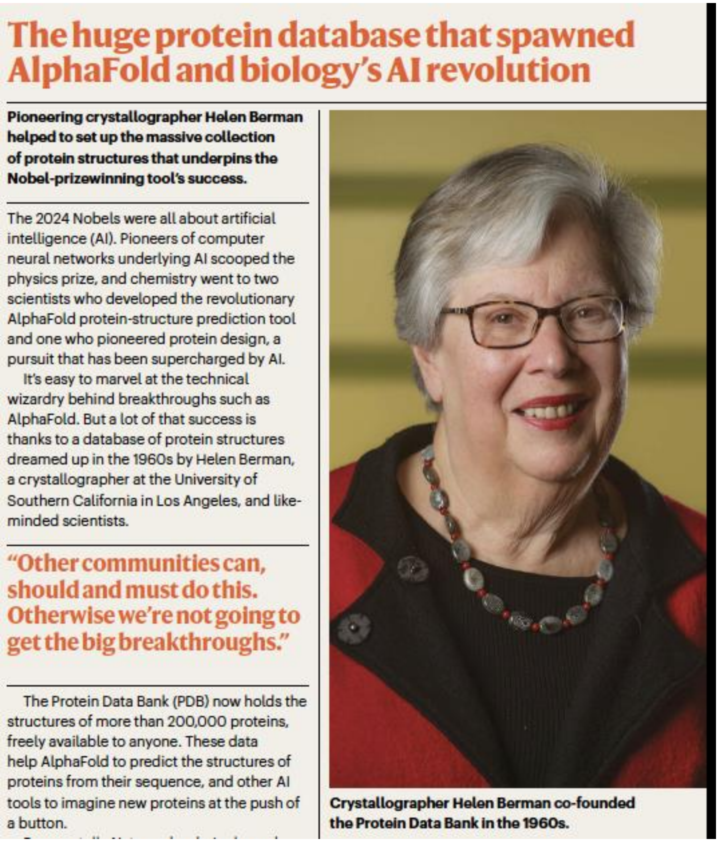
(RoseTTA) Baek, M. et al. *Science* **373,** (2021) PMID: 34282049

Excellent perspective & overview: “The impact of AlphaFold2 one year on.” *Nature Methods* (2022). PMID: 35017725

How experimental data spawned AlphaFold

After winning a share of this year's Nobel Prize in Chemistry for co-developing AlphaFold, theoretical chemist John Jumper recognized that it success came in no small part thanks to resources such as the Protein Data Bank (PDB), a freely available repository of more than 200,000 protein structures determined using methods including X-ray crystallography and cryo-electron microscopy. "It's humbling every time we train [AlphaFold] on years of effort. Each data point is years of effort from someone," he said. The PDB was dreamed up in the 1960s by crystallographer Helen Berman and like-minded scientists. Berman tells *Nature* about how the PDB has jump-started discovery, starting from the early days of a handful of structures recorded on punchcards.

https://www.nature.com/articles/d41586-024-03423-0



The huge protein database that spawned AlphaFold and biology's AI revolution

Pioneering crystallographer Helen Berman helped to set up the massive collection of protein structures that underpins the Nobel-prizewinning tool's success.

The 2024 Nobels were all about artificial intelligence (AI). Pioneers of computer neural networks underlying AI scooped the physics prize, and chemistry went to two scientists who developed the revolutionary AlphaFold protein-structure prediction tool and one who pioneered protein design, a pursuit that has been supercharged by AI.

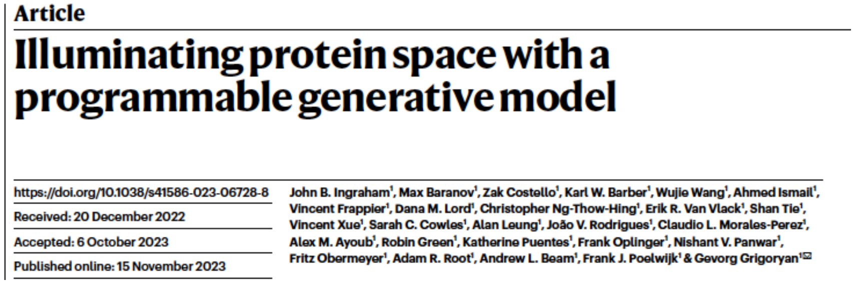
It's easy to marvel at the technical wizardry behind breakthroughs such as AlphaFold. But a lot of that success is thanks to a database of protein structures dreamed up in the 1960s by Helen Berman, a crystallographer at the University of Southern California in Los Angeles, and like-minded scientists.

"Other communities can, should and must do this. Otherwise we're not going to get the big breakthroughs."

The Protein Data Bank (PDB) now holds the structures of more than 200,000 proteins, freely available to anyone. These data help AlphaFold to predict the structures of proteins from their sequence, and other AI tools to imagine new proteins at the push of a button.

Crystallographer Helen Berman co-founded the Protein Data Bank in the 1960s.

De novo design of protein structure and function



Article

Illuminating protein space with a programmable generative model

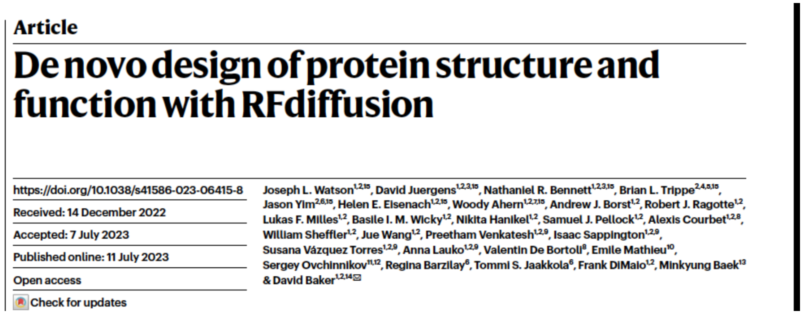
https://doi.org/10.1038/s41586-023-06728-8

Received: 20 December 2022

Accepted: 6 October 2023

Published online: 15 November 2023

John B. Ingraham[1], Max Baranov[1], Zak Costello[1], Karl W. Barber[1], Wujie Wang[1], Ahmed Ismail[1], Vincent Frappier[1], Dana M. Lord[1], Christopher Ng-Thow-Hing[1], Erik R. Van Vlack[1], Shan Tie[1], Vincent Xue[1], Sarah C. Cowles[1], Alan Leung[1], João V. Rodrigues[1], Claudio L. Morales-Perez[1], Alex M. Ayoub[1], Robin Green[1], Katherine Puentes[1], Frank Oplinger[1], Nishant V. Panwar[1], Fritz Obermeyer[1], Adam R. Root[1], Andrew L. Beam[1], Frank J. Poelwijk[1] & Gevorg Grigoryan[1] ✉



Article

De novo design of protein structure and function with RFdiffusion

https://doi.org/10.1038/s41586-023-06415-8

Received: 14 December 2022

Accepted: 7 July 2023

Published online: 11 July 2023

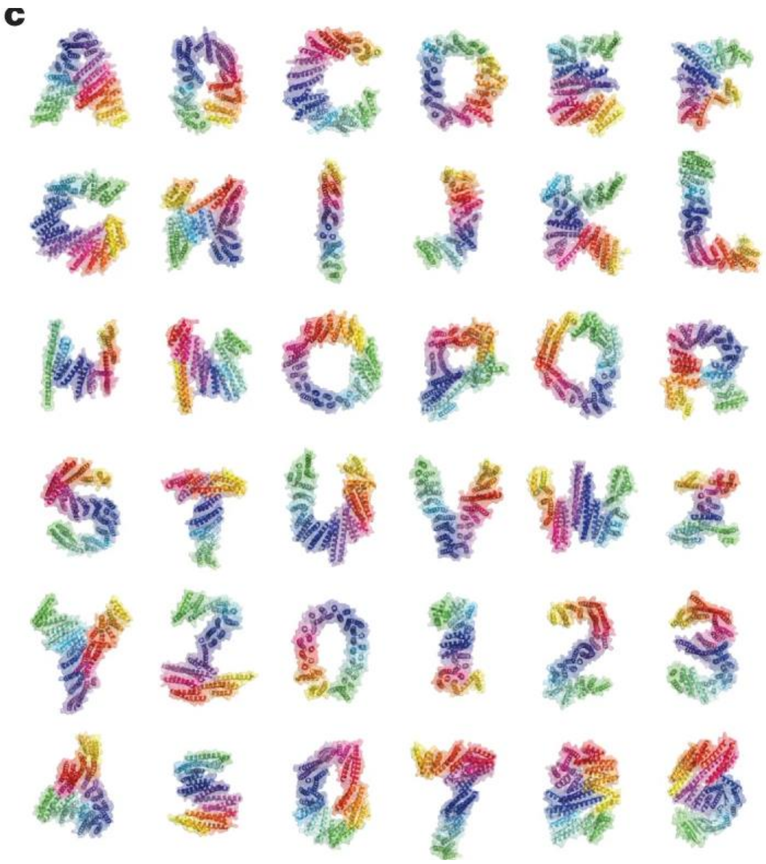
Open access

Check for updates

Joseph L. Watson[1,2,15], David Juergens[1,2,3,15], Nathaniel R. Bennett[1,2,3,15], Brian L. Trippe[2,4,5,15], Jason Yim[2,6,15], Helen E. Eisenach[1,2,15], Woody Ahern[1,2,7,15], Andrew J. Borst[1,2], Robert J. Ragotte[1,2], Lukas F. Milles[1,2], Basile I. M. Wicky[1,2], Nikita Hanikel[1,2], Samuel J. Pellock[1,2], Alexis Courbet[1,2,8], William Sheffler[1,2], Jue Wang[1,2], Preetham Venkatesh[1,2,9], Isaac Sappington[1,2,9], Susana Vázquez Torres[1,2,9], Anna Lauko[1,2,9], Valentin De Bortoli[8], Emile Mathieu[10], Sergey Ovchinnikov[11,12], Regina Barzilay[6], Tommi S. Jaakkola[6], Frank DiMaio[1,2], Minkyung Baek[13] & David Baker[1,2,14] ✉

Generating Novel, Designable, and Diverse Protein Structures by Equivariantly Diffusing Oriented Residue Clouds

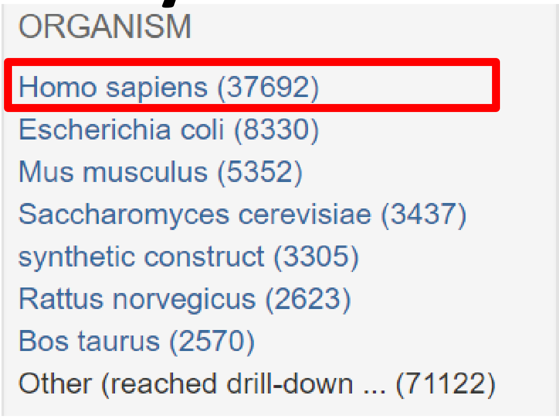
Yeqing Lin[1 2] Mohammed AlQuraishi[1 2]



1. Watson, J. L. et al. Nature https://doi.org/10.1038/s41586-023-06415-8 (2023).
2. Lin, Y. & AlQuraishi, M. Preprint at https://arxiv.org/abs/2301.12485 (2023).
3. Ingraham, J. et al. Preprint at bioRxiv https://doi.org/10.1101/2022.12.01.518682 (2022).

AI deep-learning-based methods have revealed a more complete picture of protein structure

X-ray



ORGANISM

Homo sapiens (37692)
Escherichia coli (8330)
Mus musculus (5352)
Saccharomyces cerevisiae (3437)
synthetic construct (3305)
Rattus norvegicus (2623)
Bos taurus (2570)
Other (reached drill-down ... (71122)

AlphaFold

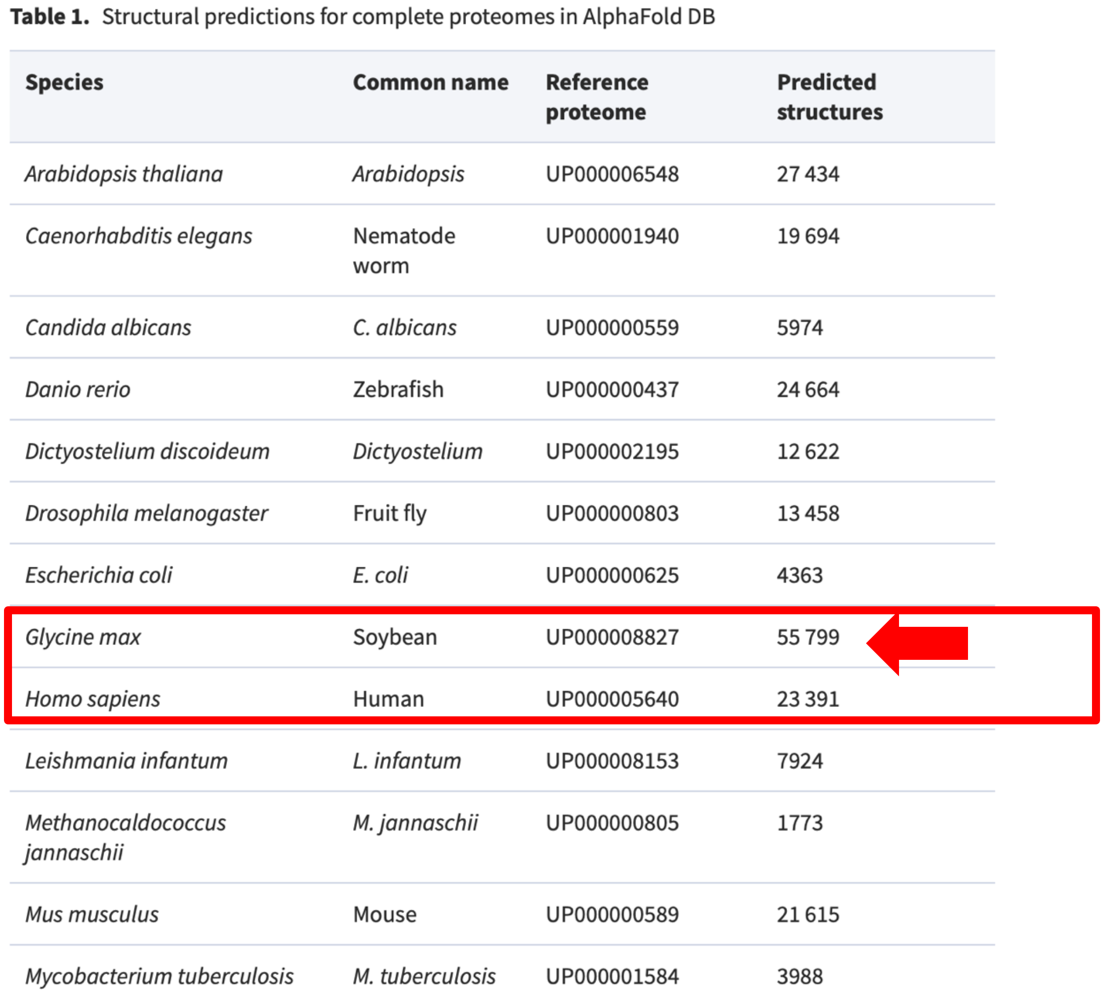


Table 1. Structural predictions for complete proteomes in AlphaFold DB

Species	Common name	Reference proteome	Predicted structures
Arabidopsis thaliana	*Arabidopsis*	UP000006548	27 434
Caenorhabditis elegans	Nematode worm	UP000001940	19 694
Candida albicans	*C. albicans*	UP000000559	5974
Danio rerio	Zebrafish	UP000000437	24 664
Dictyostelium discoideum	*Dictyostelium*	UP000002195	12 622
Drosophila melanogaster	Fruit fly	UP000000803	13 458
Escherichia coli	*E. coli*	UP000000625	4363
Glycine max	Soybean	UP000008827	55 799
Homo sapiens	Human	UP000005640	23 391
Leishmania infantum	*L. infantum*	UP000008153	7924
Methanocaldococcus jannaschii	*M. jannaschii*	UP000000805	1773
Mus musculus	Mouse	UP000000589	21 615
Mycobacterium tuberculosis	*M. tuberculosis*	UP000001584	3988

https://alphafold.ebi.ac.uk

AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Varadi M, et al. Nucleic Acids Res. 2022 PMID: 34791371