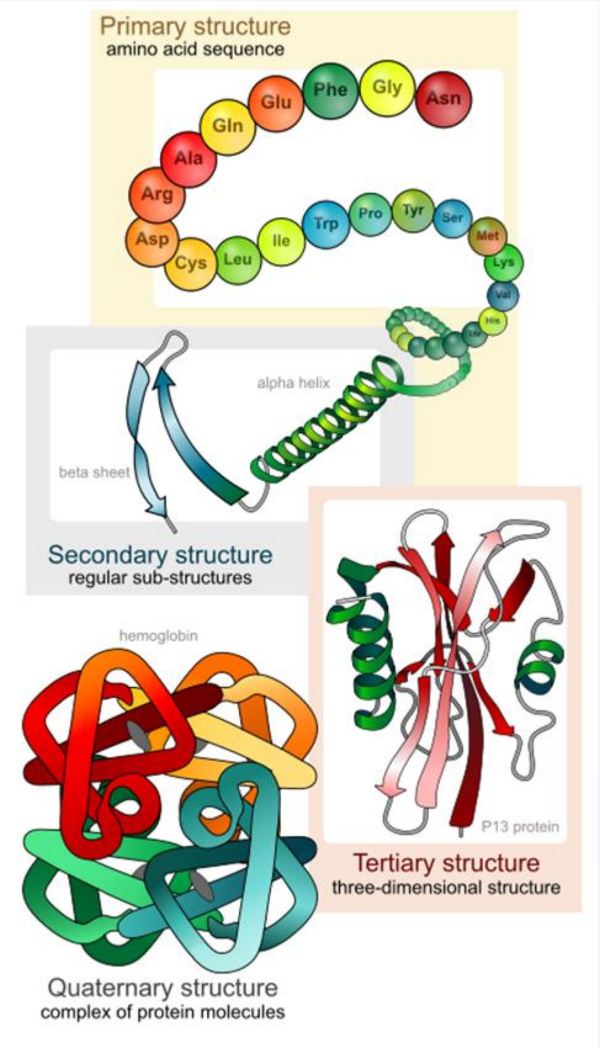
Biological DataBases

II. Protein Structure Databases

IIa. Protein Data Bank (PDB)

Overview of Protein Structures



Primary Structure

Secondary Structure

Tertiary Structure

Quarternary Structure

https://chem.libretexts.org/Bookshelves/Introductory_Chemistry/Introduction_to_Organic_and_Biochemistry_(Malik)/07%3A_Proteins/7.07%3A_Summary_of_protein_structure_levels

Introduction to PDB

- The Protein Data Bank (PDB) is a global archive of 3D structures of biological macromolecules.

- Established in 1971, managed by the Worldwide PDB (wwPDB) consortium.

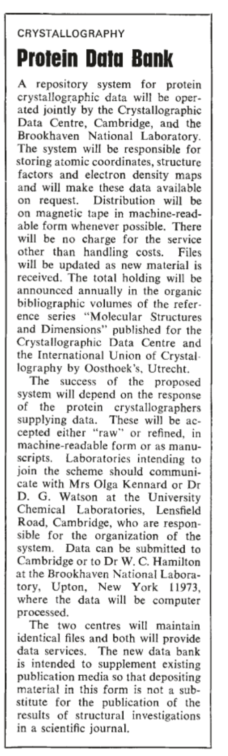
Importance of PDB

- Essential for Structural Biology and Bioinformatics.

- Provides structural information for proteins, DNA, RNA, and complexes.

History of PDB

- Founded in 1971 with only seven structures. Now contains over 250,000 structures.



CRYSTALLOGRAPHY

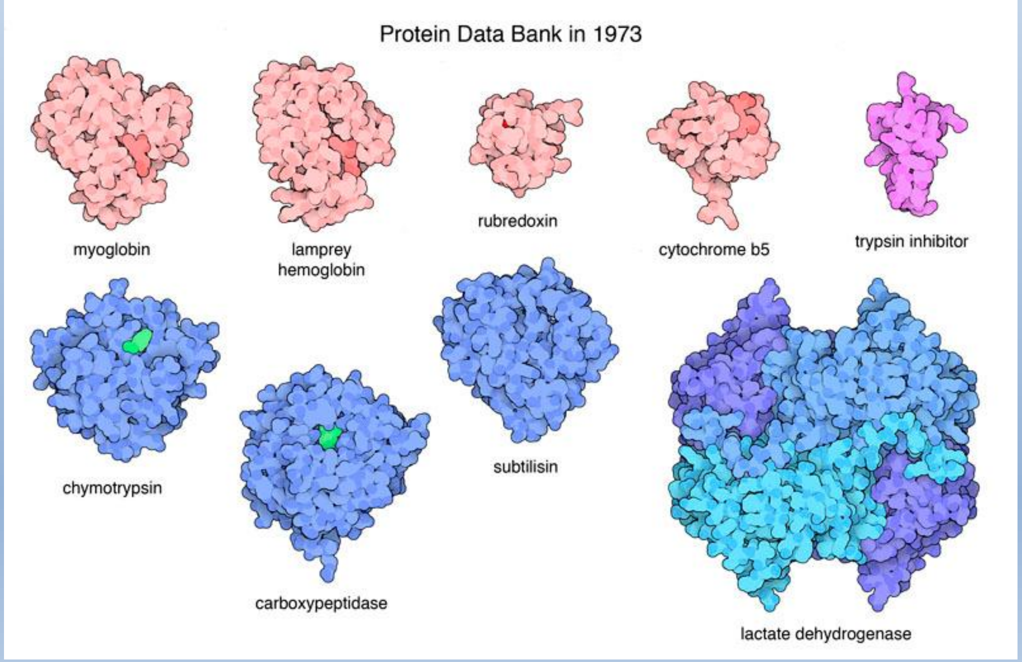
Protein Data Bank

A repository system for protein crystallographic data will be operated jointly by the Crystallographic Data Centre, Cambridge, and the Brookhaven National Laboratory. The system will be responsible for storing atomic coordinates, structure factors and electron density maps and will make these data available on request. Distribution will be on magnetic tape in machine-readable form whenever possible. There will be no charge for the service other than handling costs. Files will be updated as new material is received. The total holding will be announced annually in the organic bibliographic volumes of the reference series "Molecular Structures and Dimensions" published for the Crystallographic Data Centre and the International Union of Crystallography by Oosthoek's, Utrecht.

The success of the proposed system will depend on the response of the protein crystallographers supplying data. These will be accepted either "raw" or refined, in machine-readable form or as manuscripts. Laboratories intending to join the scheme should communicate with Mrs Olga Kennard or Dr D. G. Watson at the University Chemical Laboratories, Lensfield Road, Cambridge, who are responsible for the organization of the system. Data can be submitted to Cambridge or to Dr W. C. Hamilton at the Brookhaven National Laboratory, Upton, New York 11973, where the data will be computer processed.

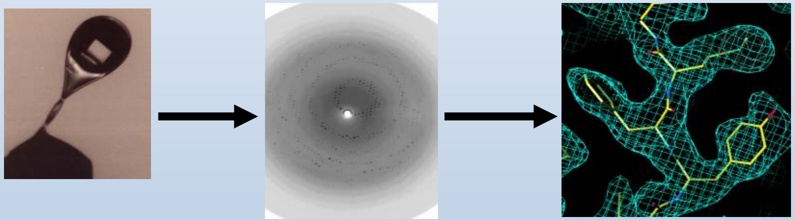
The two centres will maintain identical files and both will provide data services. The new data bank is intended to supplement existing publication media so that depositing material in this form is not a substitute for the publication of the results of structural investigations in a scientific journal.

Nature New Biology volume 233, page 223 (1971)

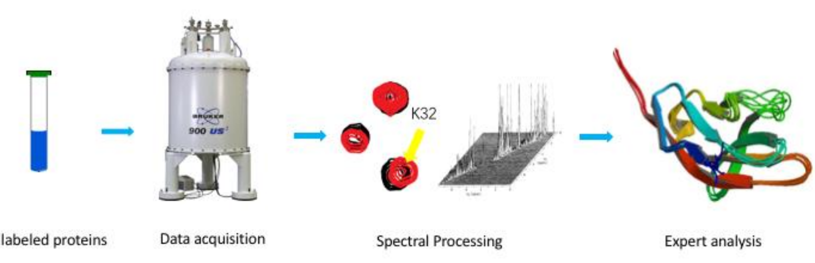


Methods of Protein Structure Determination

- X-ray Crystallography

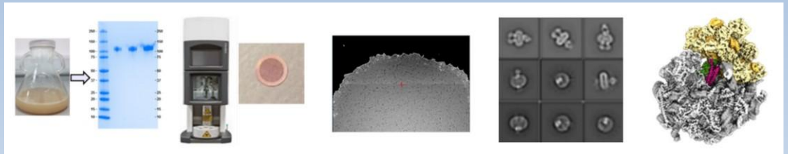


- Nuclear Magnetic Resonance Spectroscopy (NMR)



https://www.creative-biostructure.com/nmr-platform_67.htm

- Cryo-Electron Microscopy (Cryo-EM)



https://www.evotec.com/structural-biology/cryo-em

Steps in Structure Determination

- Sample Preparation

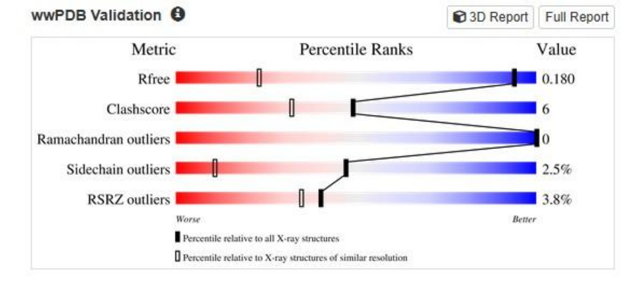
(e.g. Molecular cloning, Protein Purification, Protein Crystallization etc)

- Data Collection

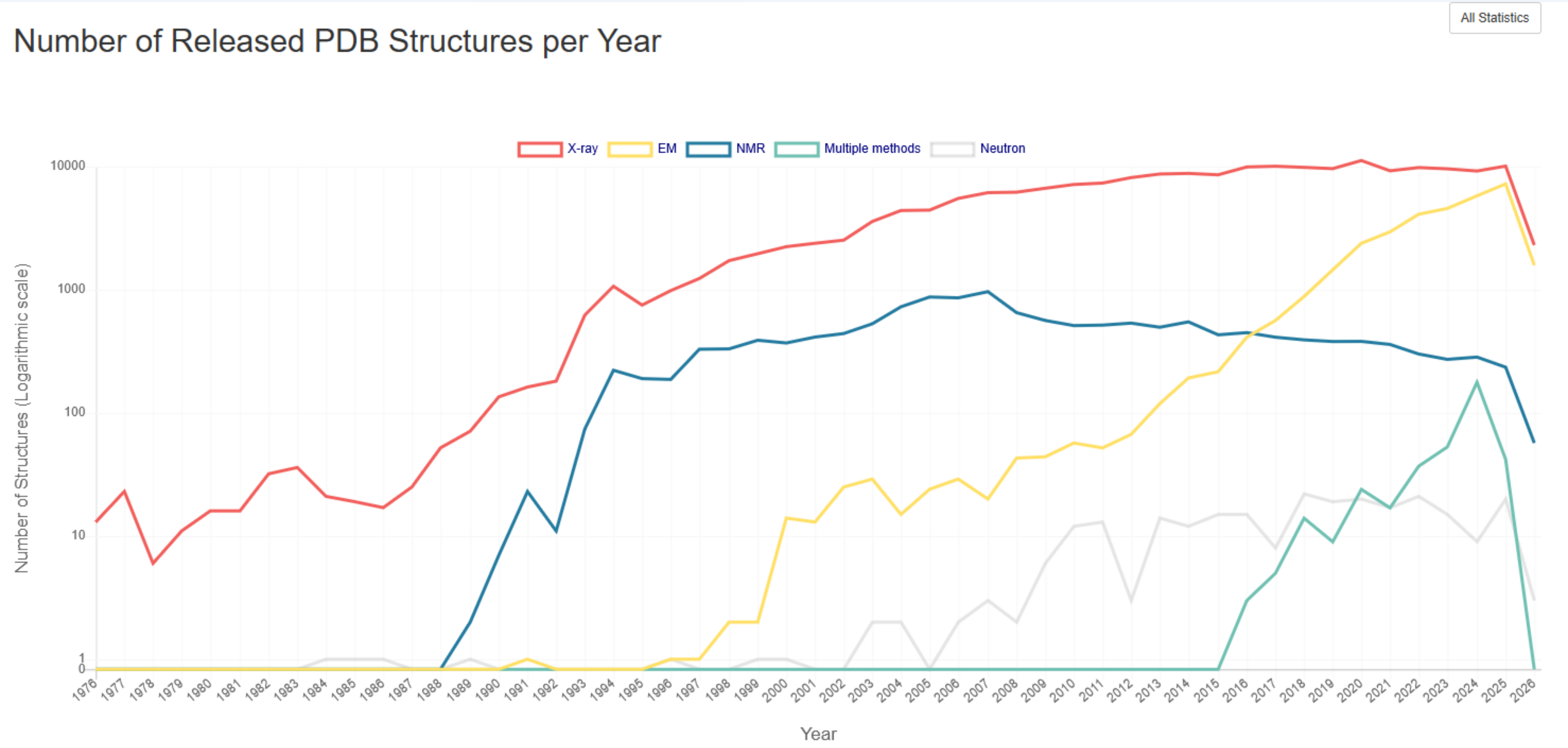
(e.g. Synchrotron, rotating anode, NMR instrument, electron microscope)

- Structure Refinement (Computational work)

- Validation & Submission



Metric	Value
Rfree	0.180
Clashscore	6
Ramachandran outliers	0
Sidechain outliers	2.5%
RSRZ outliers	3.8%



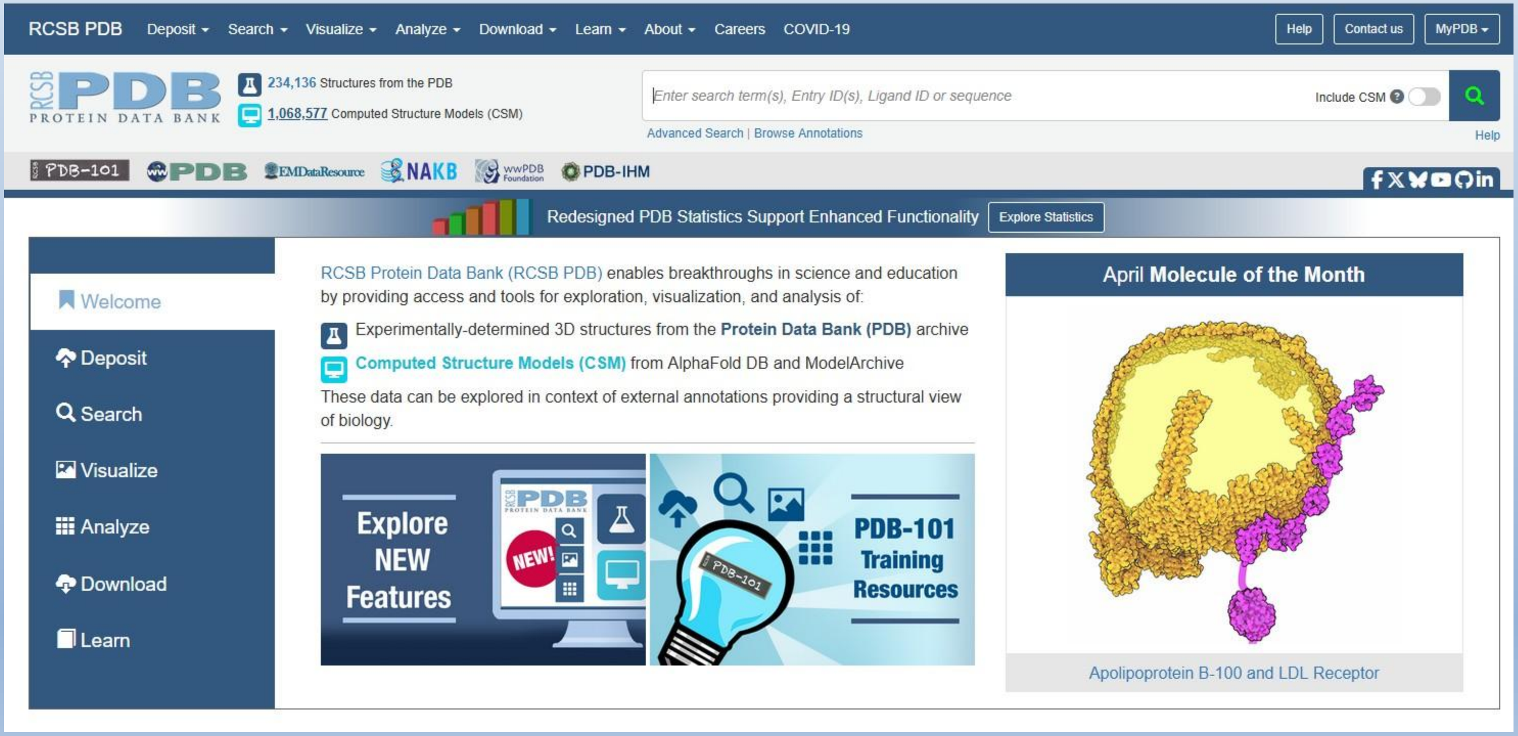
Number of Released PDB Structures per Year

- 202,825
- 32,901
- 14,722

~80,000 Structures of Human Sequences

Accessing PDB

- Website: www.rcsb.org



234,136 Structures from the PDB

1,068,577 Computed Structure Models (CSM)

Include CSM

Advanced Search | Browse Annotations

Redesigned PDB Statistics Support Enhanced Functionality

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.

Apolipoprotein B-100 and LDL Receptor

- Search by PDB ID, molecule name, keyword, or sequence.

Components of PDB

- PDB files are flat files that contain atomic coordinates, connectivity, and metadata.

[PDB File in PDB Format]()

[PDB File in CIF Format]()

Interpreting PDB Files (1HFR)

```
ATOM      1  N   ILE A   1      20.154  38.292   6.204  1.00 12.46           N
ATOM      2  CA  ILE A   1      21.273  38.823   5.380  1.00 11.92           C
ATOM      3  C   ILE A   1      22.356  37.857   5.019  1.00 11.41           C
...
HETATM 1121  C1  MTX A 200      15.421  42.873  10.422  1.00 19.67           C
HETATM 1122  N2  MTX A 200      16.634  42.948  10.931  1.00 19.55           N
```

ATOM / HETATM	ATOM = standard residue, HETATM = ligand/cofactor
ILE	Isoleucine (residue name)
A	Chain A
1	Residue number
20.154 38.292 6.204	X, Y, Z coordinates in Ångströms
1.00	Occupancy
12.46	B-factor (thermal motion)
MTX	Ligand: Methotrexate

BESIDES THAT INFORMATION:

HEADER / TITLE / COMPND

Metadata about the protein.

CONECT

Bonds between atoms—important for ligands or small molecules.

SEQRES

The full amino acid sequence (not necessarily seen in the 3D structure).

HELIX / SHEET

Information about secondary structure elements (alpha-helices, beta-sheets)

Quality Assessment of PDB Structures

- Resolution

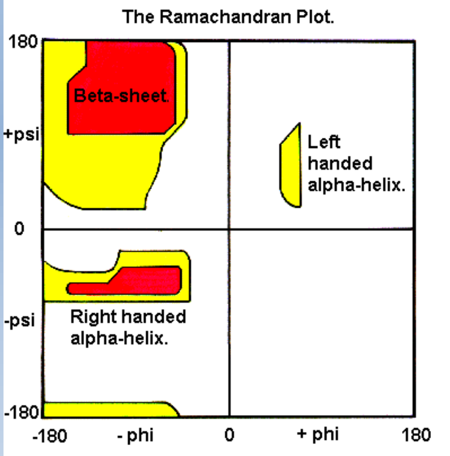
The distance corresponding to the smallest observable feature: if two objects are closer than this distance, they appear as one combined blob rather than two separate objects.

Typically, 1 – 3 Å

- R-factor

A measure of the disagreement between the crystallographic model and the experimental X-ray diffraction data - lower the R value lower is the disagreement or better is the agreement.

- Ramachandran Plot.



Applications of PDB in Research

- **Structural Biology**

(e.g. Understanding Protein Structure & Function, Protein-Ligand Interaction Studies, Comparative Structural Analysis, Structural Basis of Mutations, Macromolecular Complex Analysis……….)

- **Evolutionary Studies**

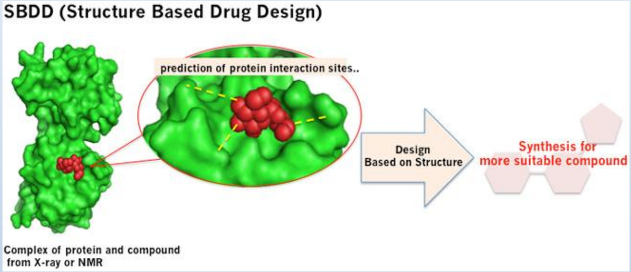
(e.g. Structural Conservation Across Species, Detecting Distant Homology……….)

- **Molecular Mechanisms**

(e.g. Enzyme Catalysis, Signal Transduction ……..)

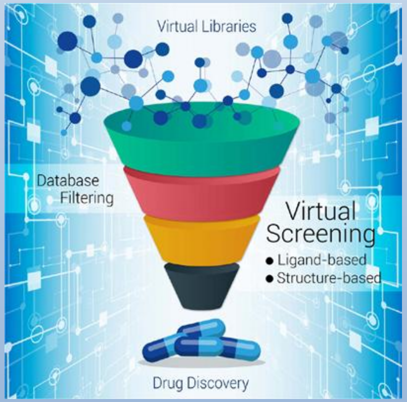
Applications of PDB in Drug Discovery

- Structure-Based Drug Design

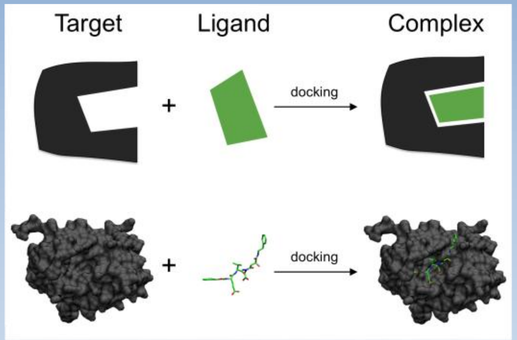


https://www.pharm.hokudai.ac.jp/soyaku/en/laboratory_01.html

- Virtual Screening

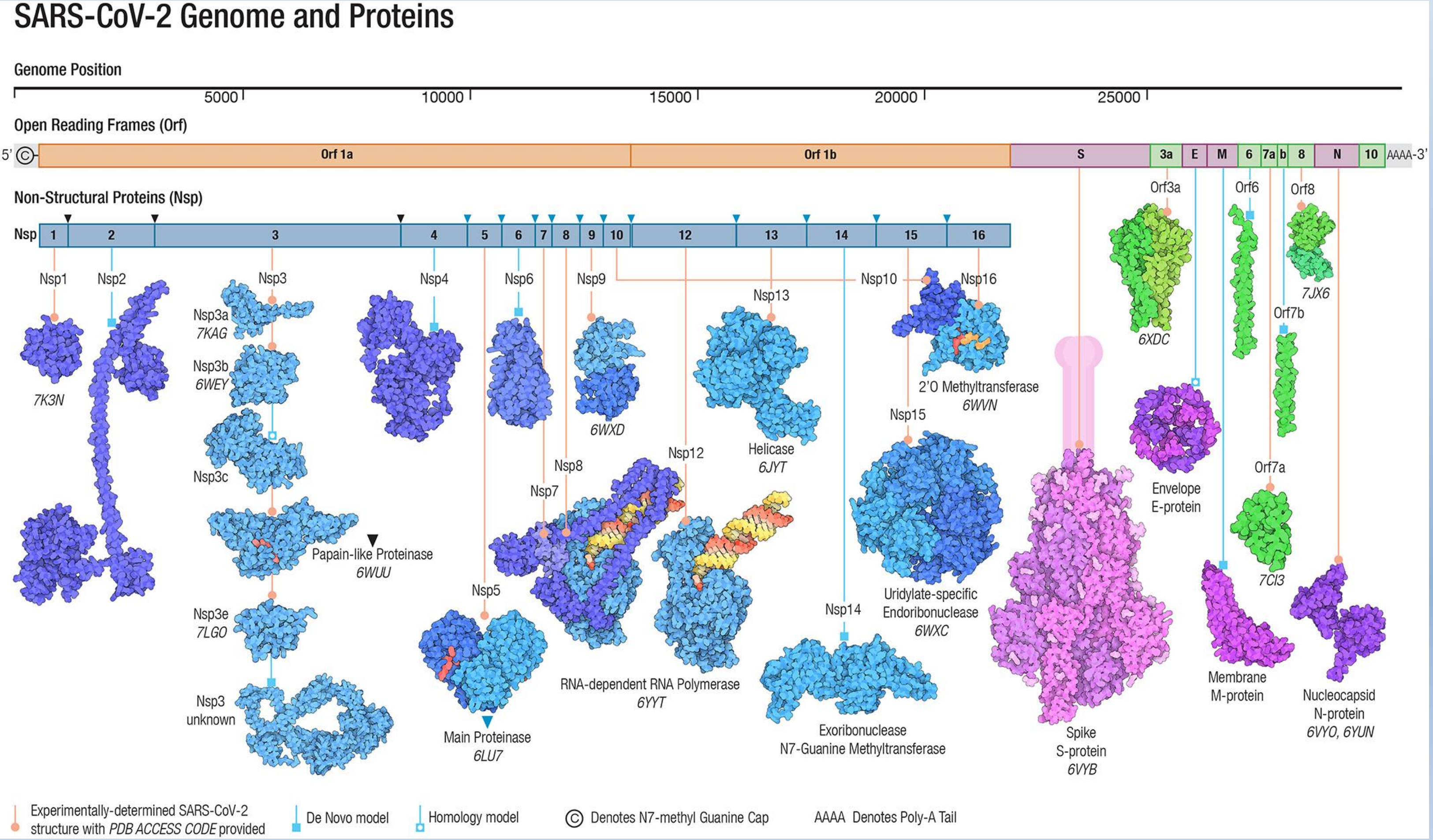


- Molecular Docking



Case Study – COVID-19 Proteins

- PDB ID: 6VYB – Structure of SARS-CoV-2 spike protein.



https://doi.org/10.1002/prot.26250

PDB in Biotechnology

Protein engineering

Process of designing or modifying proteins to improve their function, stability, or other properties.

How PDB data can help:

1. Structure => Function
2. Active Site Analysis
3. Mutation Planning
4. Protein-Protein & Protein-Ligand Interactions
5. Templates for Homology Modeling

Bioinformatics and PDB

- Homology Modeling

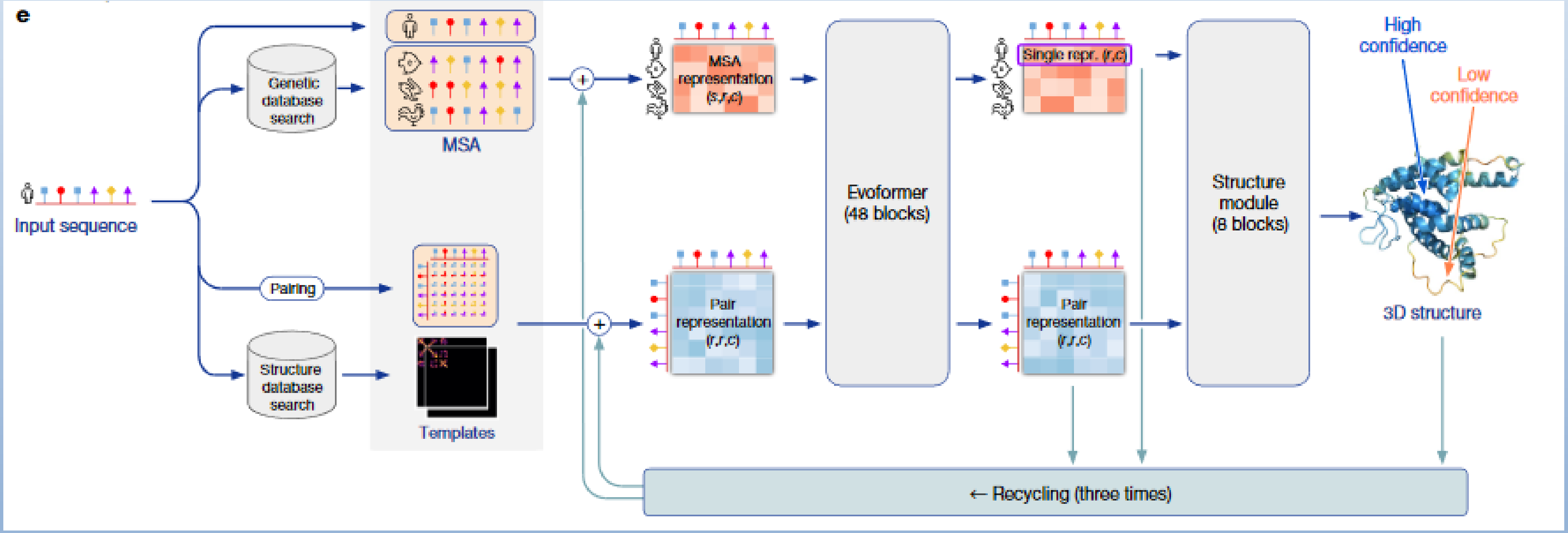
- Protein-Protein Interactions

- Evolutionary Relationships.

AI & Machine Learning in Protein Structure Prediction

- AlphaFold and Deep Learning applications

AlphaFold2 Pipeline



Challenges in Structural Biology

Depend on the method applied…

- Difficulty in crystallizing some proteins
(Transmembrane Proteins e.g. GPCRs)

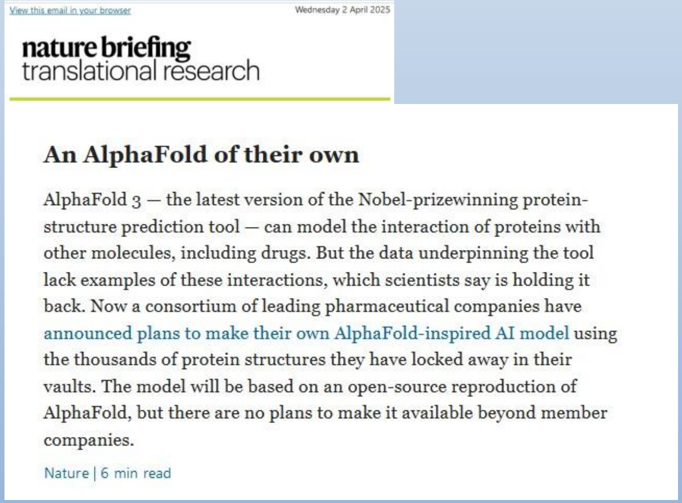
- Computational limitations

The Future of PDB

- Increasing use of AI
- Cryo-EM advancements

PDB Policies and Data Sharing

- Open access model



View this email in your browser

Wednesday 2 April 2025

nature briefing
translational research

An AlphaFold of their own

AlphaFold 3 — the latest version of the Nobel-prizewinning protein-structure prediction tool — can model the interaction of proteins with other molecules, including drugs. But the data underpinning the tool lack examples of these interactions, which scientists say is holding it back. Now a consortium of leading pharmaceutical companies have announced plans to make their own AlphaFold-inspired AI model using the thousands of protein structures they have locked away in their vaults. The model will be based on an open-source reproduction of AlphaFold, but there are no plans to make it available beyond member companies.

Nature | 6 min read

On the contrary big Pharma……

- Deposition requirements for new structures

Student & Researcher Resources

- Online tutorials, free tools, and training.

Train

Guide to Understanding PDB Data

Training Courses

Education Corner

PDB and Data Archiving Curriculum

Train: All Training Courses

Unlock Rapid Analyses Across the Whole PDB Using BinaryCIF

Teaching Enzymology with the Protein Data Bank: From Pandemic to Paxlovid (2024)

Understanding PDB Validation: Which experimental structures should I rely on?

A Deep Dive into Computed Structure Model Exploration at RCSB.org

Visualize Biomolecular Structures with Mol*: From Atoms to Movies

Teaching Enzymology with the Protein Data Bank: From Pandemic to Paxlovid (2023)

Leveraging RCSB PDB APIs for Bioinformatics Analyses and Machine Learning

Use PDB data to their full extent: Understanding PDBx/mmCIF

Conclusion

- PDB is a crucial resource for biological and medical research
- Open access to PDB accelerates scientific discoveries

IIb. Structural Classification Databases

Classify protein structures (structural domains) based on their evolutionary and structural relationships.

CATH: Class, Architecture, Topology, Homologous superfamily.

SCOP: Structural Classification of Proteins.

Applications:

- Protein function prediction.
- Evolutionary biology studies.
- Drug design and protein engineering.

CATH Database Overview

Developed by: UCL (University College London).

Hierarchy:

Class: Secondary structure composition (e.g., all-alpha, all-beta).

Architecture: Overall shape of the domain.

Topology: Fold level—how secondary structures are connected.

Homologous superfamily: Evolutionarily related domains.

Features:

Hierarchical classification of domains.

Combines automated and manual curation.

SCOP Database Overview

Developed by: MRC Laboratory of Molecular Biology, Cambridge.

Hierarchy:

- **Class:** Based on secondary structure.
- **Fold:** Major structural motifs.
- **Superfamily:** Common evolutionary origin.
- **Family:** Clear evolutionary relationship.

Features:

- Entirely manually curated (original versions).
- Updated versions include **SCOPe** (extended), **SCOP2** (revised hierarchy).
- Focused on evolutionary relationships and structural similarities.