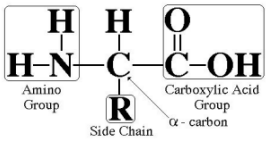
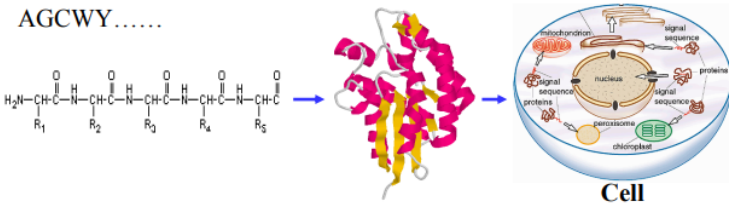
Protein Structure Modeling: Solving the Protein Folding Problem

Amino Acid Structure

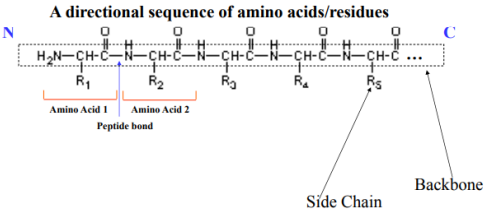




Protein Sequence – Primary Structure

- The first protein was sequenced by Frederick Sanger in 1953.
- Twice Nobel Laureate (1958, 1980) (other: Curie, Pauling, Bardeen).
- Determined the amino acid sequence of insulin and proved proteins have specific primary structure.

Protein Sequence



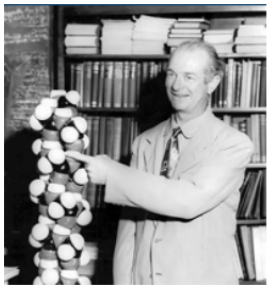
Amino Acids

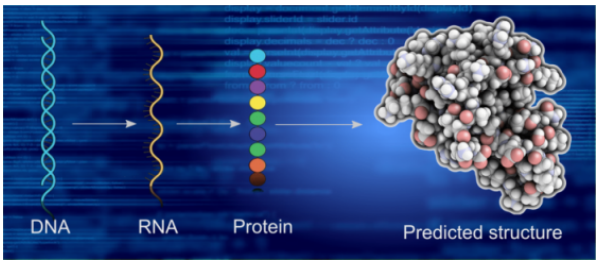
Amino acid	Abbrev.	Side chain	Hydro-phobic	Polar	Charged	Small	Tiny	Aromatic or Aliphatic	van der Waals volume	Codon	Occurrence in proteins (%)
Alanine	Ala, A	$-CH_3$	X	-	-	X	X	-	67	GCU, GCC, GCA, GCG	7.8
Cysteine	Cys, C	$-CH_2SH$	X	-	-	X	-	-	86	UGU, UGC	1.9
Aspartate	Asp, D	$-CH_2COOH$	-	X	negative	X	-	-	91	GAU, GAC	5.3
Glutamate	Glu, E	$-CH_2CH_2COOH$	-	X	negative	-	-	-	109	GAA, GAG	6.3
Phenylalanine	Phe, F	$-CH_2C_6H_5$	X	-	-	-	-	Aromatic	135	UUU, UUC	3.9
Glycine	Gly, G	-H	X	-	-	X	X	-	48	GGU, GGC, GGA, GGG	7.2
Histidine	His, H	$-CH_2-C_3H_3N_2$	-	X	positive	-	-	Aromatic	118	CAU, CAC	2.3
Isoleucine	Ile, I	$-CH(CH_3)CH_2CH_3$	X	-	-	-	-	Aliphatic	124	AUU, AUC, AUA	5.3
Lysine	Lys, K	$-(CH_2)_4NH_2$	-	X	positive	-	-	-	135	AAA, AAG	5.9
Leucine	Leu, L	$-CH_2CH(CH_3)_2$	X	-	-	-	-	Aliphatic	124	UUA, UUG, CUU, CUC, CUA, CUG	9.1
Methionine	Met, M	$-CH_2CH_2SCH_3$	X	-	-	-	-	-	124	AUG	2.3
Asparagine	Asn, N	$-CH_2CONH_2$	-	X	-	X	-	-	96	AAU, AAC	4.3
Proline	Pro, P	$-CH_2CH_2CH_2-$	X	-	-	X	-	-	90	CCU, CCC, CCA, CCG	5.2
Glutamine	Gln, Q	$-CH_2CH_2CONH_2$	-	X	-	-	-	-	114	CAA, CAG	4.2
Arginine	Arg, R	$-(CH_2)_3NH-C(NH)NH_2$	-	X	positive	-	-	-	148	CGU, CGC, CGA, CGG, AGA, AGG	5.1
Serine	Ser, S	$-CH_2OH$	-	X	-	X	X	-	73	UCU, UCC, UCA, UCG, AGU,AGC	6.8
Threonine	Thr, T	$-CH(OH)CH_3$	X	X	-	X	-	-	93	ACU, ACC, ACA, ACG	5.9
Valine	Val, V	$-CH(CH_3)_2$	X	-	-	X	-	Aliphatic	105	GUU, GUC, GUA, GUG	6.6
Tryptophan	Trp, W	$-CH_2C_8H_6N$	X	-	-	-	-	Aromatic	163	UGG	1.4
Tyrosine	Tyr, Y	$-CH_2-C_6H_4OH$	X	X	-	-	-	Aromatic	141	UAU, UAC	3.2

Hydrophilic

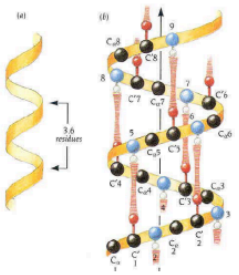
Protein Secondary Structure

- Determined by hydrogen bond patterns
- 3-Class categories: alpha-helix, beta-sheet, loop (or coil)
- First deduced by Linus Pauling et al.



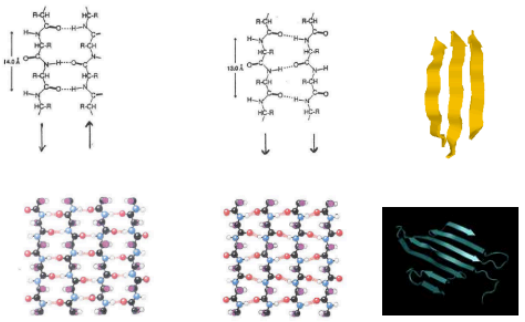


Alpha-Helix



Jurnak, 2003

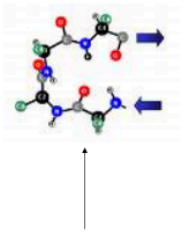
Beta-Sheet



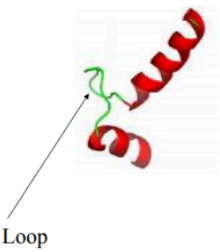
Anti-Parallel

Parallel

Non-Repetitive Secondary Structure



Beta-Turn

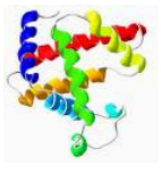


Loop

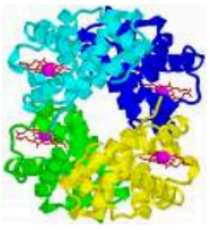
Tertiary Structure

- John Kendrew et al., Myoglobin
- Max Perutz et al., Haemoglobin
- 1962 Nobel Prize in Chemistry

Perutz Kendrew

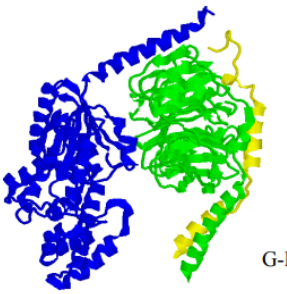


myoglobin

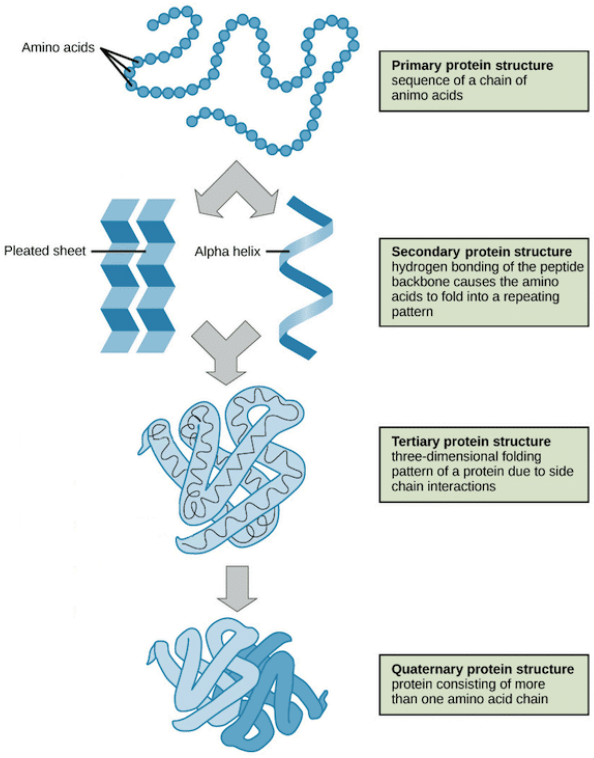


haemoglobin

Quaternary Structure: Complex



G-Protein Complex



Protein Extraction Methods:

Proteins are essential macromolecules that perform a variety of functions in the body, like DNA replication, catalyzing reactions, and providing structural support. They are studied in three main ways:

1. **In Vivo**: Studying proteins within the organism to understand how they interact.
2. **In Vitro**: Studying purified proteins in controlled lab settings to avoid interference from other factors.
3. **In Silico**: Using computer simulations to study proteins, saving time and resources.

Proteins are classified into types like **extracellular matrix proteins** (e.g., elastin, collagen) and **globular proteins** (e.g., enzymes, antibodies). Purifying proteins from other cellular components is crucial for research, whether for large-scale production (e.g., insulin) or analysis of small protein amounts.

Uses of Isolated Proteins:

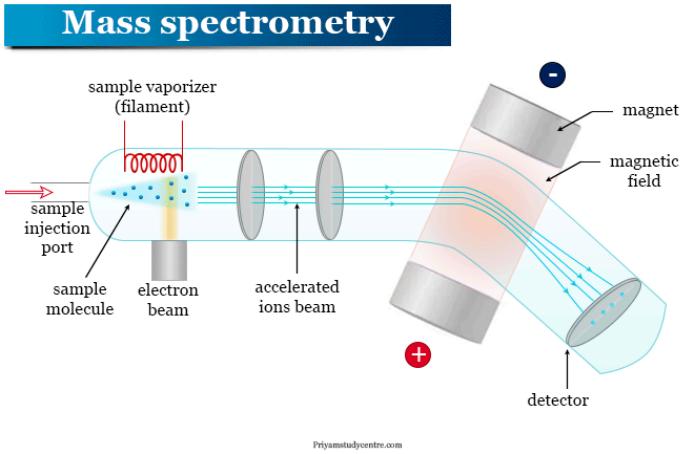
Protein extraction is widely applied in both research and industry. The purification process requires multiple steps and detection methods, including absorbance, spectrometry, and antibody-based techniques.

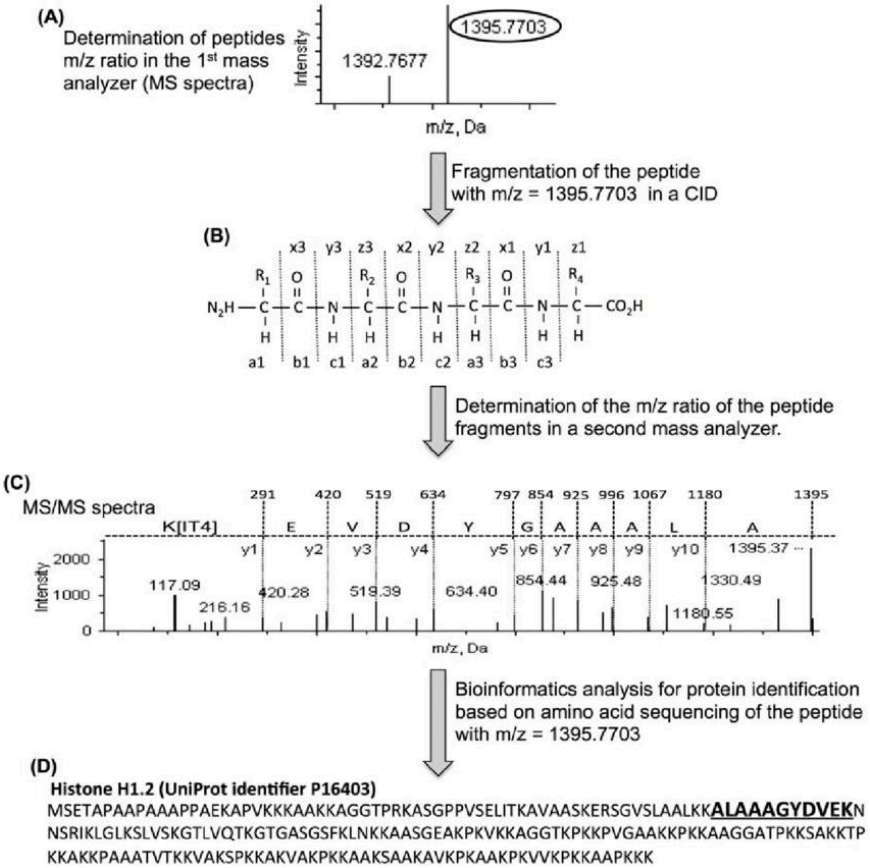
In **clinical applications**, isolated proteins can help diagnose diseases like diabetes or be used in treatments (e.g., collagen in skincare). In **research**, purified proteins enable various studies:

Immunoprecipitation (IP): Isolates proteins using antibodies.
Proteomics: Studies the entire set of proteins in an organism.
Enzyme Assays: Measure enzyme activity, including different experiment types like relaxation or transient kinetics.
Western Blot: Detects specific proteins in a sample.

Gel Electrophoresis: Separates proteins by size and charge.
Biomarkers: Used to track biological processes like disease or treatment effects.

- Mass Spectrometry: It analyzes protein fragments to determine their sequence and is useful for complex mixtures. Mass Spectrometry (MS) measures the **mass-to-charge** ratio of ions to identify and quantify molecules. Mass spectrometry is a tool that measures the mass-to-charge ratio of molecules in a sample. This helps determine the exact molecular weight of the components in the sample.





- Next-Generation Sequencing (NGS): This modern technique uses mRNA to indirectly infer protein sequences, offering high throughput.

Anfinsen's Folding Experiment

- Structure is uniquely determined by protein sequence
- Protein function is determined by protein structure

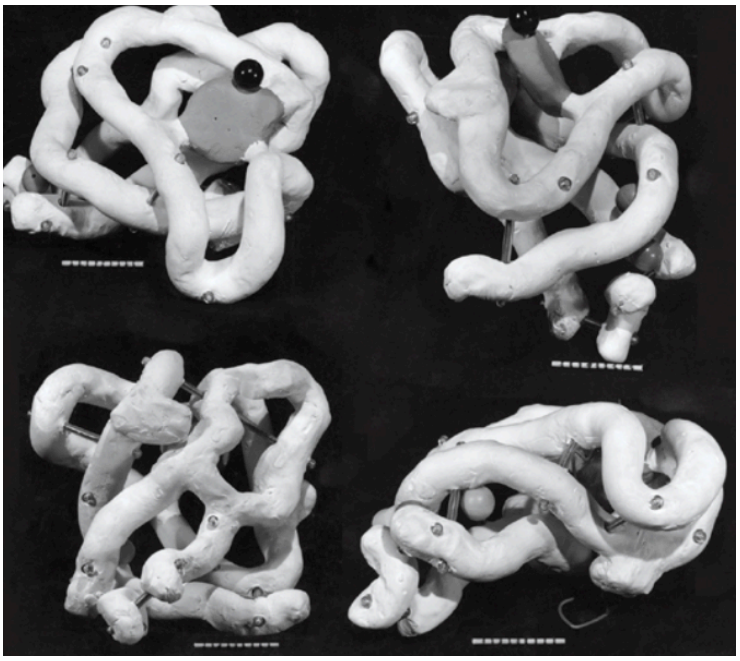
Protein Structure Determination

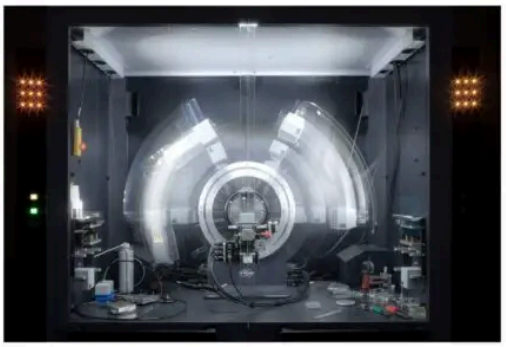
- X-ray crystallography
- Nuclear Magnetic Resonance (NMR) Spectroscopy
- X-ray: any size, accurate (1-3 Angstrom (10^{-10} m)), sometime hard to grow crystal
- NMR: small to medium size, moderate accuracy, structure in solution

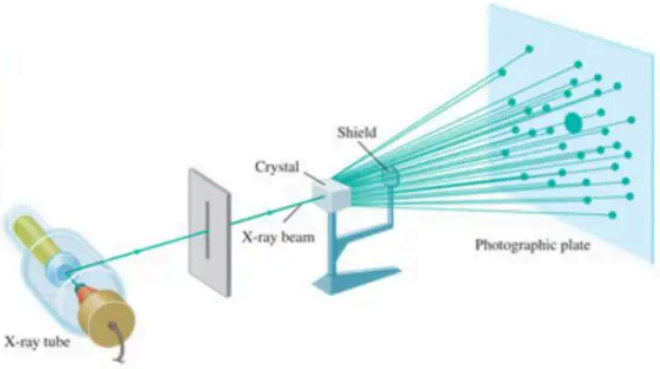
X-ray crystallography is a key method for determining protein structures, offering detailed data on atomic arrangement. It works by passing X-rays through a rotating protein crystal and analyzing the diffracted rays. The technique provides high resolution but requires high-quality crystals and large amounts of protein, often produced through recombinant methods. It is particularly effective for rigid proteins but less so for flexible ones.

The three-dimensional structure of myoglobin was first published by John Kendrew and his team in 1958. This image, recreated by the author using original figures from the Medical Research Council Laboratory of Molecular Biology, shows the low-resolution structure. The white polypeptide chains represent the protein, and the grey disc is the haem group. The three spheres indicate the positions where heavy atoms were attached: black represents mercury from p-chloro-mercuri-benzene-sulphonate, dark grey is mercury from mercury diammine, and light grey is gold from auri-chloride. The scale marks are 1 Å (angstrom) apart.

In the 1930s, William Astbury studied diffraction from biological fibers, while Dorothy Crowfoot (Hodgkin) and J.D. Bernal explored crystals of macromolecules. **Over 20 years later, in 1958, John Kendrew and his team solved the first crystal structure of myoglobin from sperm whale muscle (given below).** This was followed by Max Perutz's work on haemoglobin (1962) and David Phillips on lysozyme (1965).



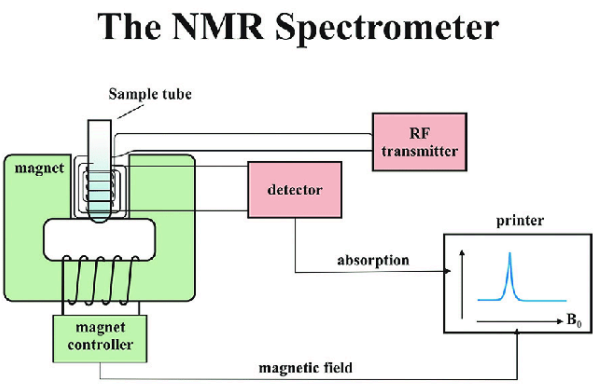


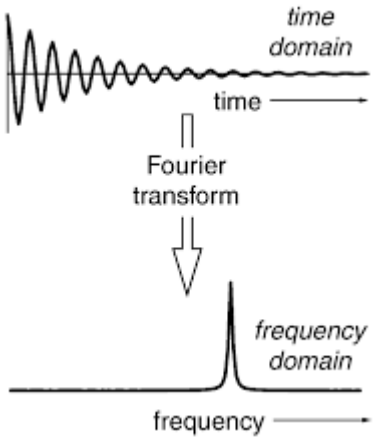


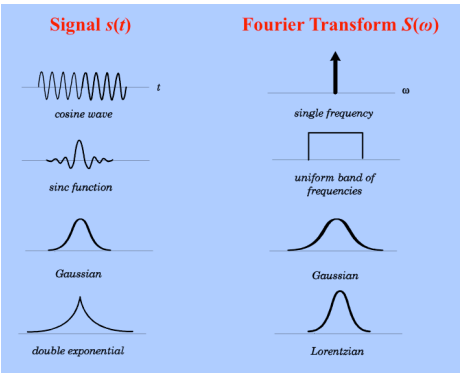
Meanwhile, in the 1930s, Rabi and colleagues demonstrated **nuclear resonance (NMR)** by applying electromagnetic radiation to molecular beams. Though NMR was theoretically possible in solids and liquids, early attempts failed due to low sensitivity and long relaxation times of the nuclei. Advances in electronics led to the first NMR spectra in condensed phases by groups led by Bloch, Pound, Purcell, and others in the mid-1940s. The chemical shift effect was observed by Knight (1949), Proctor and Yu (1950-1951), and Dickinson (1950), and spin-spin coupling was discovered by Hahn and Maxwell (1951).

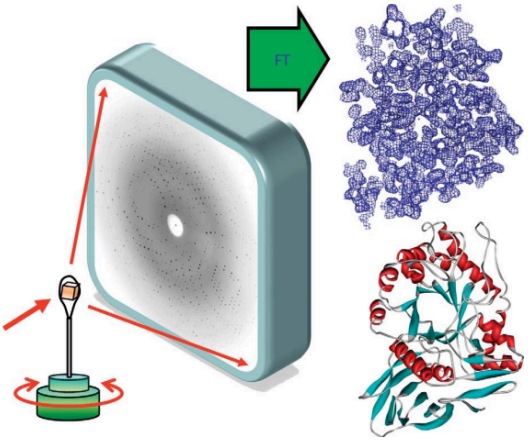
Nuclear Magnetic Resonance (NMR) Spectroscopy: This technique uses radiofrequency waves to analyze protein atoms. It requires larger quantities of stable protein at room temperature and works best for small proteins, offering high resolution, especially for flexible proteins.

Fourier transform (FT) techniques introduced by Ernst and Anderson in 1966 paved the way for two-dimensional NMR experiments, transforming NMR's role in biological systems. The first proton NMR spectrum for a protein was recorded in 1957 (Saunders et al.), and since then, NMR methods have advanced to enable routine assignments of proton resonances in proteins up to 50 kDa. One key discovery was the Overhauser Effect (OE) by Overhauser in 1953, which improved signal-to-noise ratios and led to the development of NOESY experiments used in protein structure determination. In 1985, Wüthrich and colleagues reported the first complete NMR structure of a globular protein (Williamson et al., 1985).









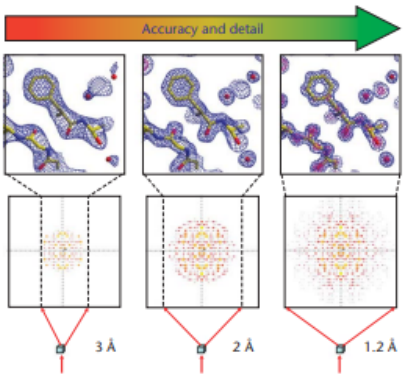
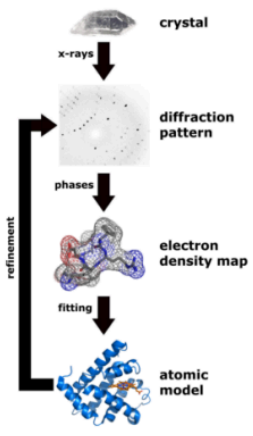


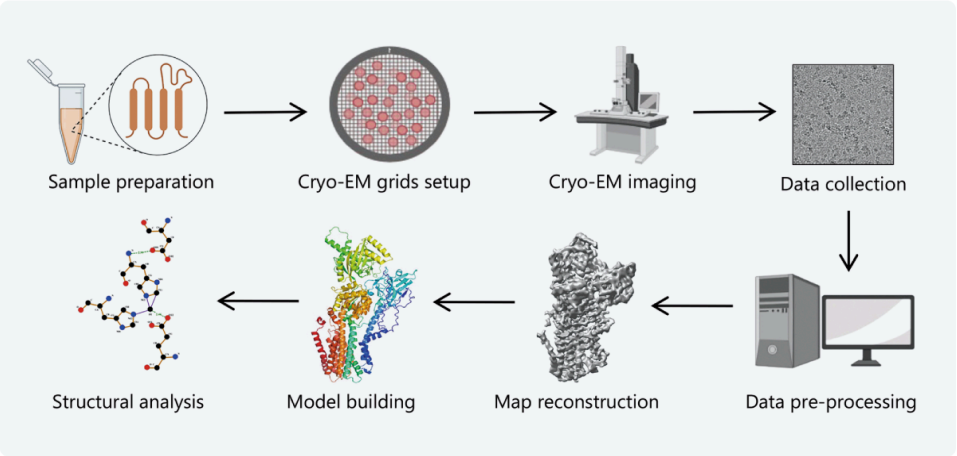
Figure 5 Data quality determines structural detail and accuracy. The qualitative relation between the extent of X-ray diffraction, the resulting amount of available diffraction data, and the quality and detail of the electron density reconstruction and protein structure model are evident from this figure: The crystals are labelled with the nominal resolution in Å as determined by the highest diffraction angle at which X-ray reflections are observed. Above the crystals is a sketch of the corresponding diffraction pattern, which contains significantly more data at higher resolution. As a consequence, both the reconstruction of the electron density (blue grid) and the resulting structure model (stick model) are much more detailed and accurate. The non-SI unit Å (10^{-8} cm or 0.1 nm = 10^{-10} m) is generally used in the crystallographic literature, simply because it is of the same order of magnitude as atomic radii (approximately 0.77 Å for carbon) or bond lengths (approximately 1.54 Å for the C–C single bond). From *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology* by Bernhard Rupp. Reproduced by permission of Garland Science/Taylor & Francis LLC © 2011.

In Fourier Transform Nuclear Magnetic Resonance (FT-NMR), the Free Induction Decay (FID) is the signal collected over time after a radiofrequency pulse. This signal is then converted into a frequency-based spectrum using a Fourier Transform.



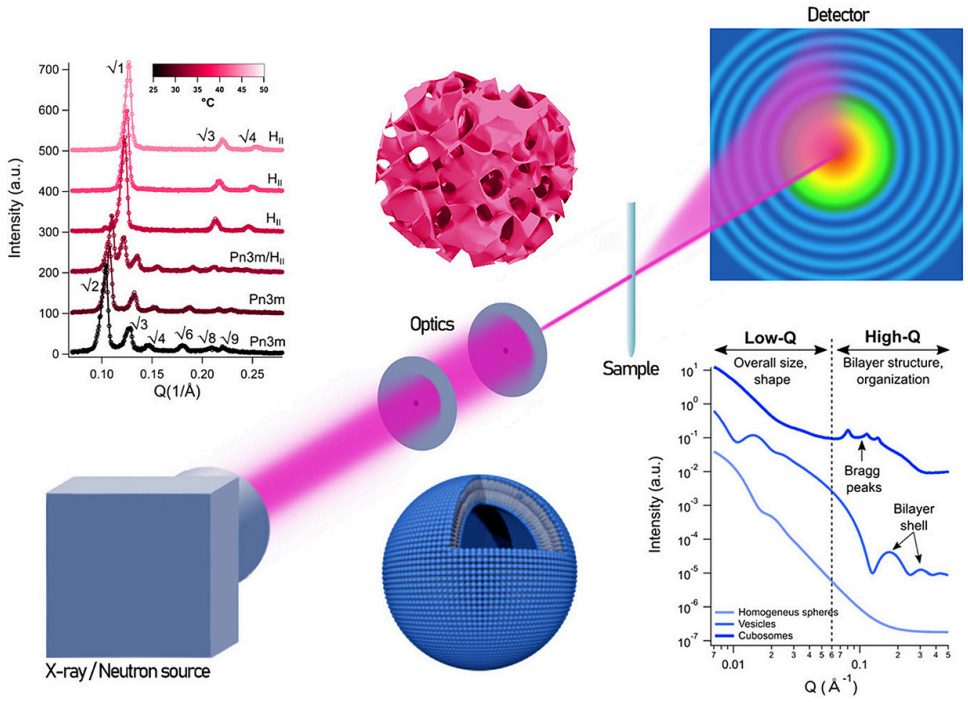
Pacific Northwest National Laboratory's high magnetic field (800 MHz 18.8 T) NMR spectrometer being loaded with a sample.

The **Protein Data Bank now holds over 206,000 protein structures, and technological improvements have led to a "resolution revolution,"** especially in cryo-EM. **Cryo-EM is a technique used in structural biology to see the 3D shapes of biological molecules.** It involves quickly freezing the molecules in a thin layer of ice and then imaging them with an electron beam.

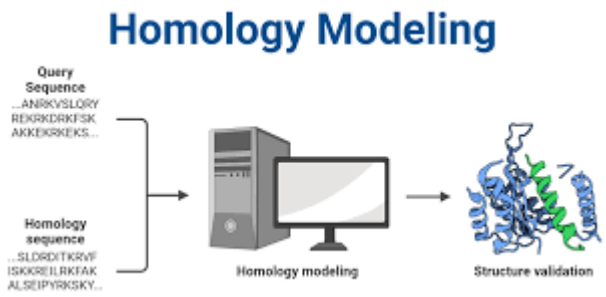


Small-Angle X-Ray Scattering (SAXS) and Neutron Scattering: These methods are useful for studying protein structures in solution when high resolution isn't necessary, allowing better control over experimental conditions.

Lipid-based nanoparticles (LNPs), which vary in structure from small vesicles to more complex forms, are increasingly used to deliver drugs, vaccines, and nutrients. Small-angle scattering techniques using X-rays (SAXS) or neutrons (SANS) are valuable tools for studying the structure, behavior, and interactions of these particles at different scales, from nano to molecular.



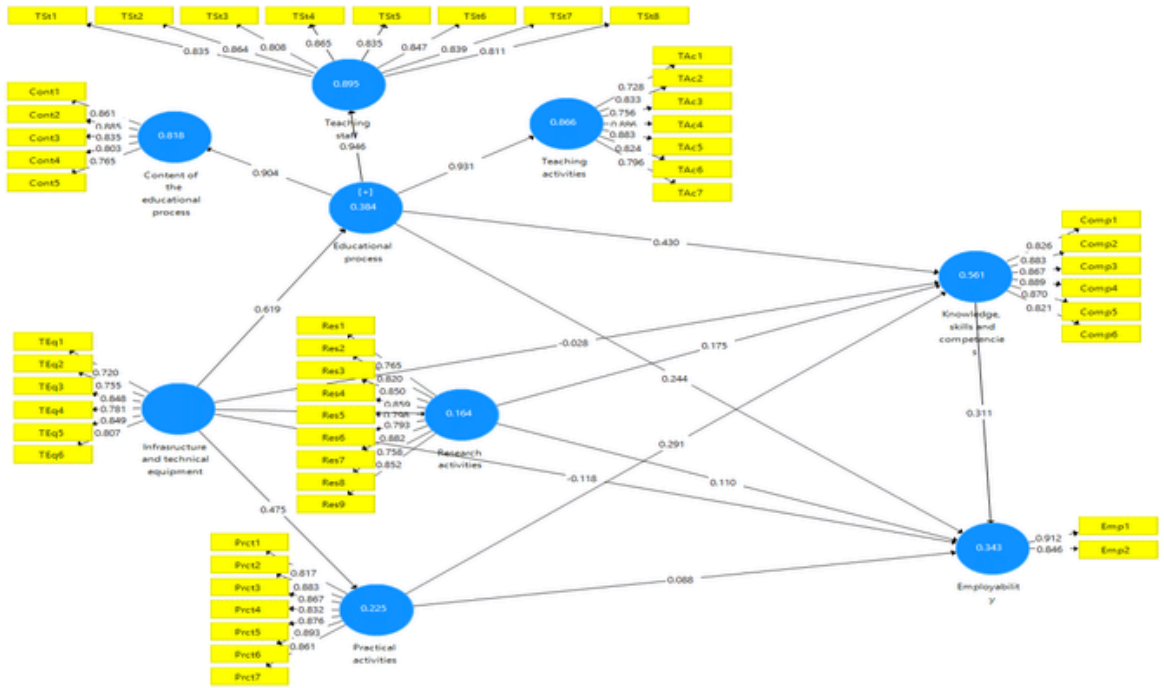
Homology Modeling: This technique creates a 3D protein model based on a known similar protein, relying on sequence similarities.



Partial Structural Study Methods: These include ultracentrifugation, mass spectrometry, and fluorescence spectrometry, often used alongside other techniques to gain further insights into protein structure.

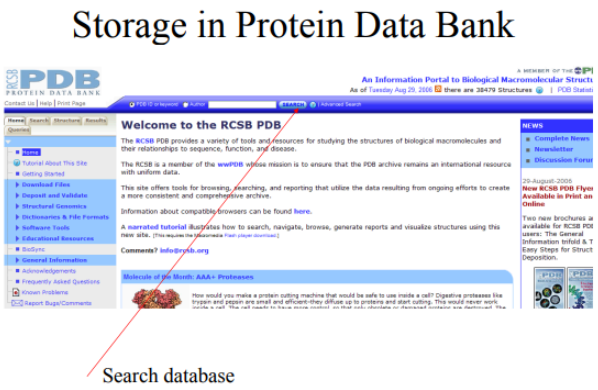
Partial Least Squares Structural Equation Modeling (PLS-SEM) is a simple method used to identify factors linked to diseases. Example: we looked at the relationship between high-sensitivity C-reactive protein (hs-CRP), blood-related inflammation markers, and the risk of cardiovascular disease (CVD). We followed 7,362 healthy participants aged 35-65 years and tracked their health over 10 years. By the end of the study, 1,022 individuals developed CVD.

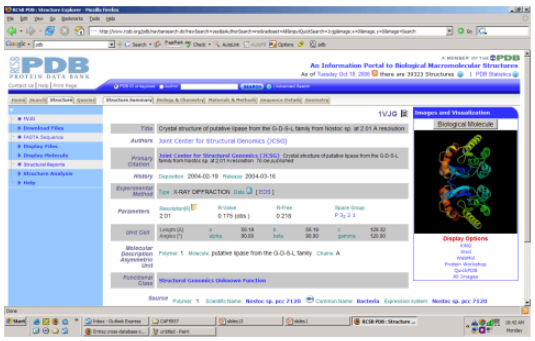
Using PLS-SEM, we built a model to predict CVD risk factors. The results showed that age was the biggest factor affecting CVD risk. Each year of age increases the risk by 0.166. Hs-CRP was the second most significant factor, raising the risk by 0.042 for each increase. The study also found a strong link between red cell distribution width (RDW) and CVD. Several other factors, like hemoglobin, neutrophils, and certain ratios, indirectly affected CVD risk through their impact on hs-CRP, even after accounting for age, sex, and socioeconomic factors. Overall, the key risk factors for CVD were age, hs-CRP, and RDW.



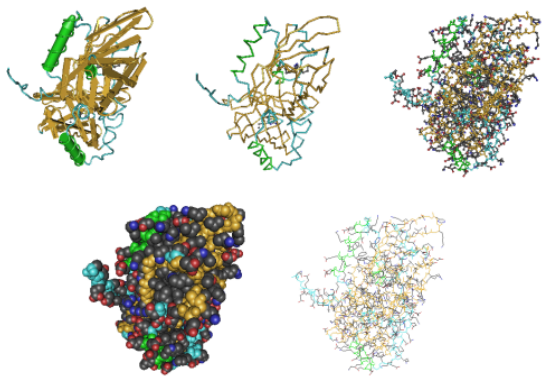
As of January 2012, the Protein Data Bank (PDB) contained about 80,000 entries, with 70,000 from X-ray crystallography and 10,000 from NMR spectroscopy. NMR is primarily used for smaller proteins (under 50 kDa), while X-ray crystallography is used for larger proteins (over 35 kDa).

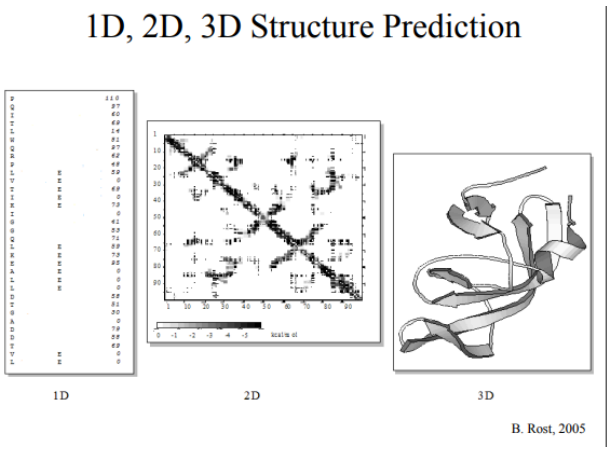
The discovery of protein structures led to the field of protein folding, where researchers aim to predict how a protein's amino acid sequence determines its 3D shape and how it folds. This "protein folding problem" has **two main challenges: predicting a protein's structure from its sequence, and understanding how it folds.** Early researchers like Levinthal and Ptitsyn suggested that folding might follow a stepwise process, with rapid formation of secondary structures like alpha-helices and beta-sheets. Modern studies use advanced technologies like recombinant DNA, protein engineering, and computer simulations to analyze folding at the atomic level. Recombinant DNA, in particular, allowed for the study of proteins that were previously difficult to obtain in large quantities, enabling better understanding of protein folding.

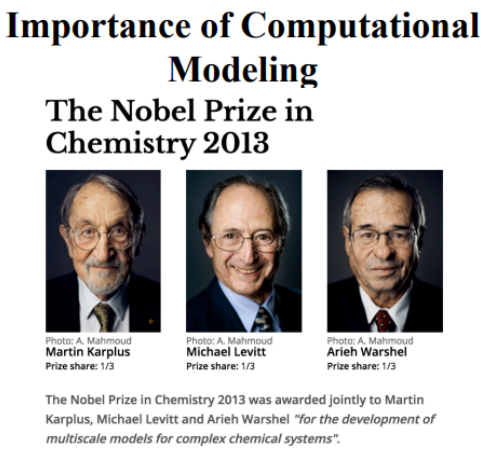




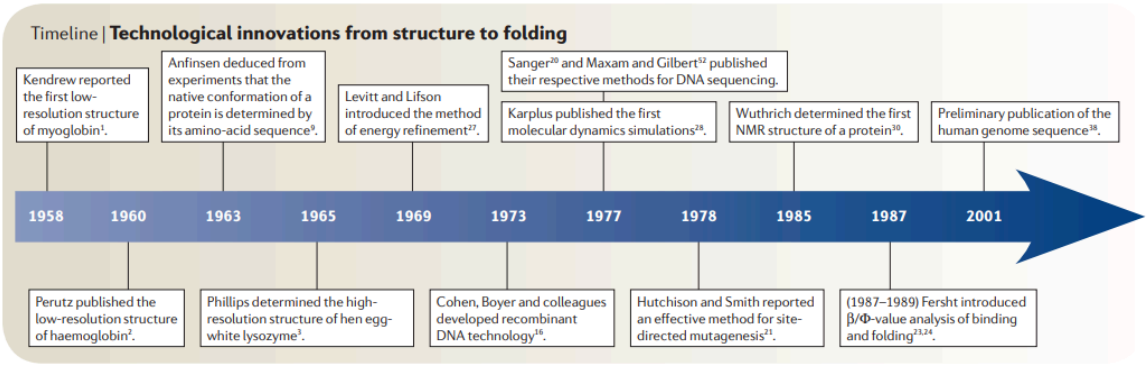

```
SEQRES   1 A   21  GLY ILE VAL GLU GLN CYS CYS THR SER ILE CYS SER LEU
SEQRES   2 A   21  TYR GLN LEU GLU ASN TYR CYS ASN
SEQRES   1 B   29  PHE VAL ASN GLN HIS LEU CYS GLY SER HIS LEU VAL GLU
SEQRES   2 B   29  ALA LEU TYR LEU VAL CYS GLY GLU ARG GLY PHE PHE TYR
SEQRES   3 B   29  THR PRO LYS
FORMUL   3  HOH   *31(H2 O1)
HELIX    1   1 GLY A    1  CYS A    7  1                                   7
HELIX    2   2 SER A   12  ASN A   18  1                                   7
HELIX    3   3 GLY B    8  GLY B   20  1                                  13
HELIX    4   4 GLU B   21  GLY B   23  5                                   3
SSBOND   1 CYS A    6    CYS A   11                          1555   1555
SSBOND   2 CYS A    7    CYS B    7                          1555   1555
SSBOND   3 CYS A   20    CYS B   19                          1555   1555
CRYST1   78.608   78.608   78.608  90.00  90.00  90.00 I 21 3       24
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      0.012721  0.000000  0.000000        0.00000
SCALE2      0.000000  0.012721  0.000000        0.00000
SCALE3      0.000000  0.000000  0.012721        0.00000
ATOM      1  N   GLY A   1      45.324  26.807  11.863  1.00 24.82           N
ATOM      2  CA  GLY A   1      45.123  27.787  12.967  1.00 24.93           C
ATOM      3  C   GLY A   1      43.756  27.627  13.605  1.00 25.16           C
ATOM      4  O   GLY A   1      43.107  26.591  13.438  1.00 25.00           O
ATOM      5  N   ILE A   2      43.313  28.661  14.323  1.00 25.21           N
ATOM      6  CA  ILE A   2      42.050  28.622  15.065  1.00 25.39           C
ATOM      7  C   ILE A   2      40.818  28.303  14.200  1.00 25.69           C
ATOM      8  O   ILE A   2      39.935  27.565  14.635  1.00 25.56           O
ATOM      9  CB  ILE A   2      41.816  29.917  15.917  1.00 25.39           C
```







https://www.rcsb.org/



RMSD has often been used to measure the quality of reproduction of a known (i.e. crystallographic) binding pose by a computational method, such as docking. In this case a low RMSD with respect to the true binding pose, is good (ideally less than 1.5Angstrom, or even better, less than 1 Angstrom). This represents good reproduction of the correct pose. Beware reading much significance into relative values of RMSD's greater than 3. A binding pose with RMSD 4 Angstrom is not better than one of 6 Angstrom. They are both equally poor (i.e. wrong)

Proteins are crucial for life, and understanding their structure helps us understand how they function. While around 100,000 protein structures have been determined, **this is only a tiny fraction of the billions of known protein sequences**. The challenge is that determining a single protein structure takes months or years of work. To address this, accurate computational methods are needed for large-scale structural bioinformatics. **Predicting a protein's 3D structure from its amino acid sequence—known as the "protein folding problem"—has been a major research challenge for over 50 years.**

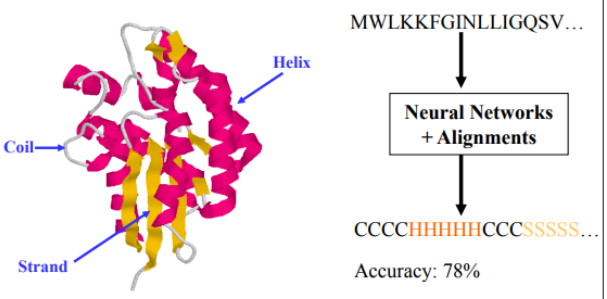
In the 1970s, Levitt and Cyrus Chothia discovered that proteins could be grouped based on their secondary structure, which led to bioinformatics methods for analyzing protein structures through similarities (homology). This requires knowledge of protein sequences, which was made possible by the Human Genome Project. Though controversial at first, the project provided invaluable data that advanced bioinformatics and personalized medicine.

Databases like the NCBI and EBI store this data and are used to predict protein structures by comparison with known ones.

The Protein Structure Initiative, a global effort, aims to map the 3D structures of all proteins by solving the structures of representative proteins and modeling others based on similarities. It's hoped that 2,000 protein structures will be enough to model all proteins. However, even if 2,000 is too few, this information will still be valuable.

Despite advances in computational power, predicting protein structures from scratch is still not fully possible. However, bioinformatics allows for high-precision models of small proteins and good models of larger, multi-domain proteins. These methods rely on accumulated experimental data and homology modeling. While predicting protein folding is still challenging, progress has been made with simulations and the study of ultra-fast-folding proteins.

1D: Secondary Structure Prediction



Pollastri et al. *Proteins*, 2002. Cheng et al.. *NAR*, 2005

How to Use Neural Network to Predict Secondary Structure

- Create a data set with input sequences (x) and output labels (secondary structures)
- Encode the input and output to neural network
- Train neural network on the dataset (training dataset)
- Test on the unseen data (test dataset) to estimate the generalization performance.

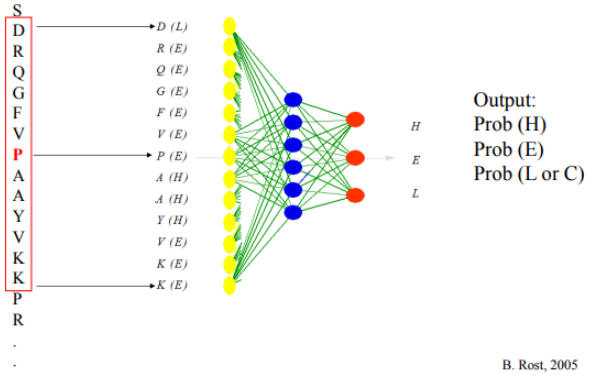
Create a Data Set

- Download proteins from Protein Data Bank
- Select high-resolution protein structures (<2.5 Angstrom, determined by X-ray crystallography)
- Remove proteins with chain-break (Ca-Ca distance > 4 angstrom)
- Remove redundancy (filter out very similar sequences using BLAST)
- Use DSSP program (Kabsch and Sander, 1983) to assign secondary structure to each residue.

Train and Test

- Use one data set as training dataset to build neural network model
- Use another data set as test dataset to evaluate the generalization performance of the model
- Sequence similarity any two sequences in test and training dataset is less than 25%.

Secondary Structure Prediction (Generation III – Neural Network)

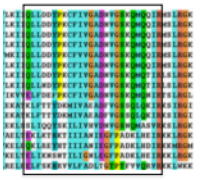


B. Rost, 2005

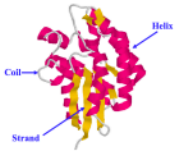
Evolutionary Information is Important

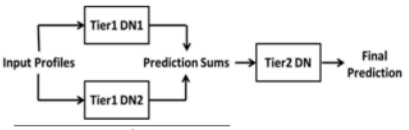
- Single sequence yields accuracy below 70%.
- Use all the sequences in the family of a query sequence can improve accuracy to 78%.
- Structure is more conserved than sequence during evolution. The conservation and variation provides key information for secondary structure prediction.

Deep Learning for Secondary Structure Prediction



Window Size = 19





Rank	Layers	Nodes[a]	Q_3(%)	Sov (%)
1	4	600	80.24	73.07
2	5	400	79.96	72.67
3	3	500	80.00	72.37
4	4	300	79.84	72.48
5	6	300	79.82	72.50
6	3	400	79.89	72.20
7	6	400	79.87	72.15
8	6	200	79.70	72.24
9	5	500	79.42	71.03
10	4	400	79.35	70.83
11	4	500	79.39	70.31
12	3	600	79.07	70.06

Spencer et al., *IEEE Transactions on Computational Biology and Bioinformatics*, 2014.

Deep Learning for Secondary Structure Prediction Project (2nd project)

- Training dataset with sequences and secondary structures (1180 sequences) and test dataset (126 sequences). (training data was created by Pollastri et al. and test data was created by Rost and Sander.) (http://calla.rnet.missouri.edu/cheng_courses/mlbioinfo/ss_train.txt(and ss_test.txt)
- Generate multiple alignments using generate_flatblast.sh in Pspro 1.2 package (http://sysbio.rnet.missouri.edu/multicom_toolbox/tools.html)

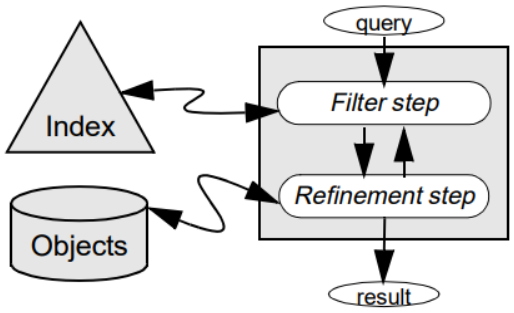


Figure 4: Multi-step similarity query processing.

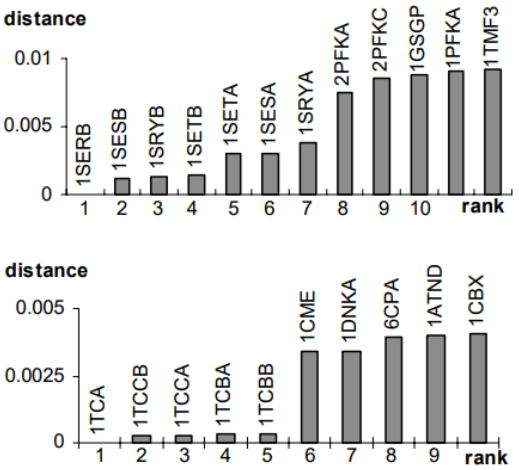
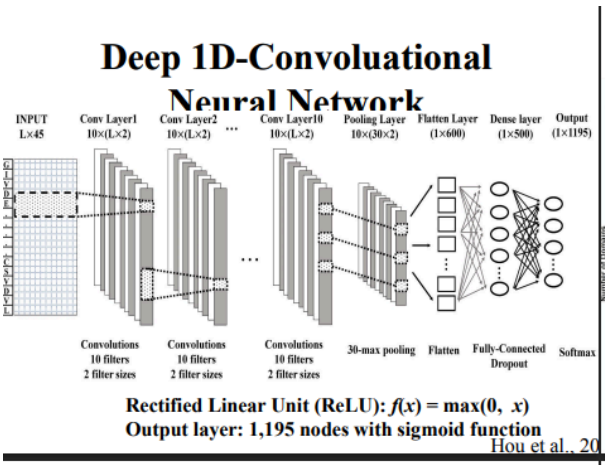
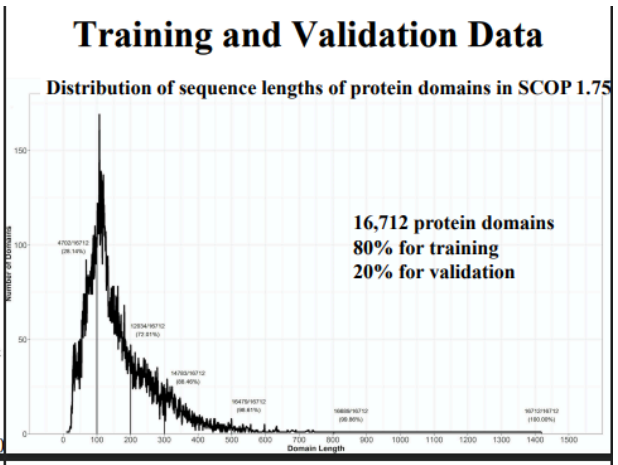
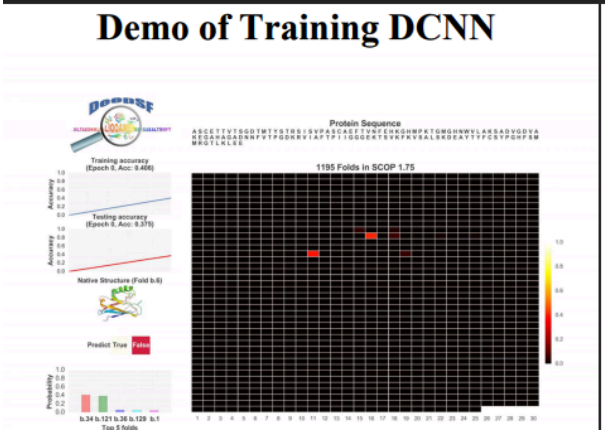
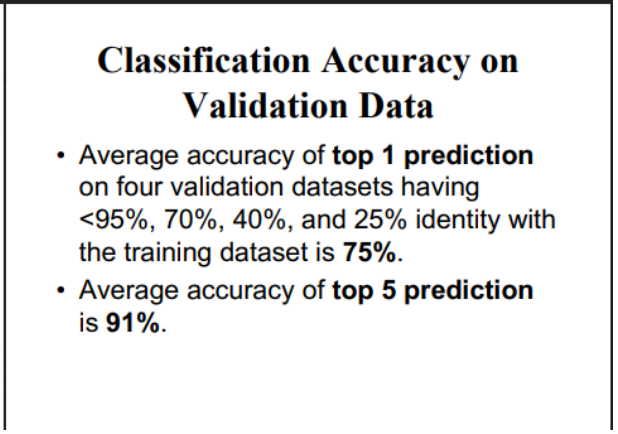


Figure 5: Similarity ranking for the seryl-tRNA synthetase 1SER-B (top) and the yeast hydrolase 1TCA (bottom) for histograms of 6 shells and 20 sectors. The diagrams depict the top nearest neighbors and their similarity distances to the query protein.









AlphaFold: A Breakthrough in Protein Structure Prediction

Article

Highly accurate protein structure prediction with AlphaFold

https://doi.org/10.1038/s41586-021-03819-2
Received: 11 May 2021
Accepted: 12 July 2021
Published online: 15 July 2021
Open access
Check for updates

John Jumper[1,4]✉, Richard Evans[1,4], Alexander Pritzel[1,4], Tim Green[1,4], Michael Figurnov[1,4], Olaf Ronneberger[1,4], Kathryn Tunyasuvunakool[1,4], Russ Bates[1,4], Augustin Žídek[1,4], Anna Potapenko[1,4], Alex Bridgland[1,4], Clemens Meyer[1,4], Simon A. A. Kohl[1,4], Andrew J. Ballard[1,4], Andrew Cowie[1,4], Bernardino Romera-Paredes[1,4], Stanislav Nikolov[1,4], Rishub Jain[1,4], Jonas Adler[1], Trevor Back[1], Stig Petersen[1], David Reiman[1], Ellen Clancy[1], Michal Zielinski[1], Martin Steinegger[2,3], Michalina Pacholska[1], Tamas Berghammer[1], Sebastian Bodenstein[1], David Silver[1], Oriol Vinyals[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Pushmeet Kohli[1] & Demis Hassabis[1,4]✉

What is AlphaFold?

AlphaFold, developed by **Google DeepMind**, is an AI system that predicts the **3D structure of proteins** with high accuracy. It has revolutionized protein research by making structure prediction faster and more accessible.

How AlphaFold2 Works

- Uses **machine learning** to predict protein structures from their amino acid sequences.

- Trained on the **Protein Data Bank (PDB)**, a database of experimentally determined structures.

- Does **not rely on templates** and can predict entirely new protein folds.

- Provides **confidence scores** (pLDDT, pTM, PAE) to assess prediction reliability.

Key Achievements

- Won the **CASP14** competition in 2020, proving its accuracy.
- Available **open-source** and via web platforms like Google Colab.
  The **AlphaFold Protein Structure Database** contains **~200 million** predicted structures.
  Widely adopted, with **15,000+ citations** and **millions of downloads** since its release.

AlphaFold2 is a **game-changer** in molecular biology, accelerating drug discovery, genetic research, and more.

The breakthrough of AlphaFold2, trained on data from the Protein Data Bank, has transformed protein structure prediction, offering accurate models and revolutionizing how structural biology is performed. However, challenges remain in predicting protein aggregation, complexes, and dynamic interactions, and further improvements are needed for full accuracy, particularly in predicting the structure of mutated proteins.

In the last 20 years, structural biology has evolved into a multimodal field, integrating various techniques to understand protein structures and their functions. Protein structure prediction algorithms, like AlphaFold2 and RoseTTAFold, have made significant strides, enabling better understanding of protein structures using deep learning. These methods, alongside tools like X-ray crystallography, cryo-EM, and NMR, provide insights into the 3D shapes of proteins and their roles in diseases and drug discovery.

AlphaFold2, developed by Google DeepMind, has revolutionized protein structure prediction using AI trained on Protein Data Bank data. It provides highly accurate 3D models of proteins but still faces challenges in predicting protein complexes, interactions, and mutations.

Structural biology now integrates AI-based methods like AlphaFold2 and RoseTTAFold with traditional techniques (X-ray crystallography, cryo-EM, NMR) to study proteins for drug discovery and disease research.

Recent advances in neural network (NN)-based protein modeling, particularly AlphaFold and RoseTTAFold, have shown great promise, especially for challenging protein families like G-protein-coupled receptors (GPCRs).

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
Oryza sativa	Asian rice	UP000059680	43 649
Plasmodium falciparum	*P. falciparum*	UP000001450	5187
Rattus norvegicus	Rat	UP000002494	21 272
Saccharomyces cerevisiae	Budding yeast	UP000002311	6040
Schizosaccharomyces pombe	Fission yeast	UP000002485	5128
Staphylococcus aureus	*S. aureus*	UP000008816	2888
Trypanosoma cruzi	*T. cruzi*	UP000002296	19 036
Zea mays	Maize	UP000007305	39 299

The Human Genome Project mapped and sequenced the entire human genome, a massive undertaking that took years and involved international collaboration. AlphaFold, an AI system developed by DeepMind, revolutionized protein structure prediction, impacting fields like drug discovery and understanding diseases. While the Human Genome Project focused on DNA sequencing, AlphaFold tackles the challenge of determining protein structures from DNA sequences.

The Human Genome Project:

Goal:

To map and sequence all human DNA, including identifying and mapping genes.

Timeline:

The project began in 1990 and was completed in 2003.

Method:

Involved sequencing DNA and analyzing the functional and physical aspects of genes.

Impact:

Provided a foundational understanding of the human genome, leading to advancements in genetics, disease research, and personalized medicine.

AlphaFold:

Goal:

To predict the three-dimensional structure of proteins from their amino acid sequence.

Method:

Uses artificial intelligence and deep learning to achieve unprecedented accuracy in protein structure prediction.

Timeline:

AlphaFold 2 emerged in 2020, significantly advancing protein structure prediction.

Impact:

Accelerates research across various fields, including drug discovery, understanding fundamental biological processes, and developing new therapies.

Accessibility:

The AlphaFold Protein Structure Database provides freely accessible predictions for over 200 million proteins.

Key Differences and Connections:

Focus:

Human Genome Project focused on DNA, while AlphaFold focuses on proteins.

Methodology:

Human Genome Project relied on traditional sequencing technologies, while AlphaFold uses AI and deep learning.

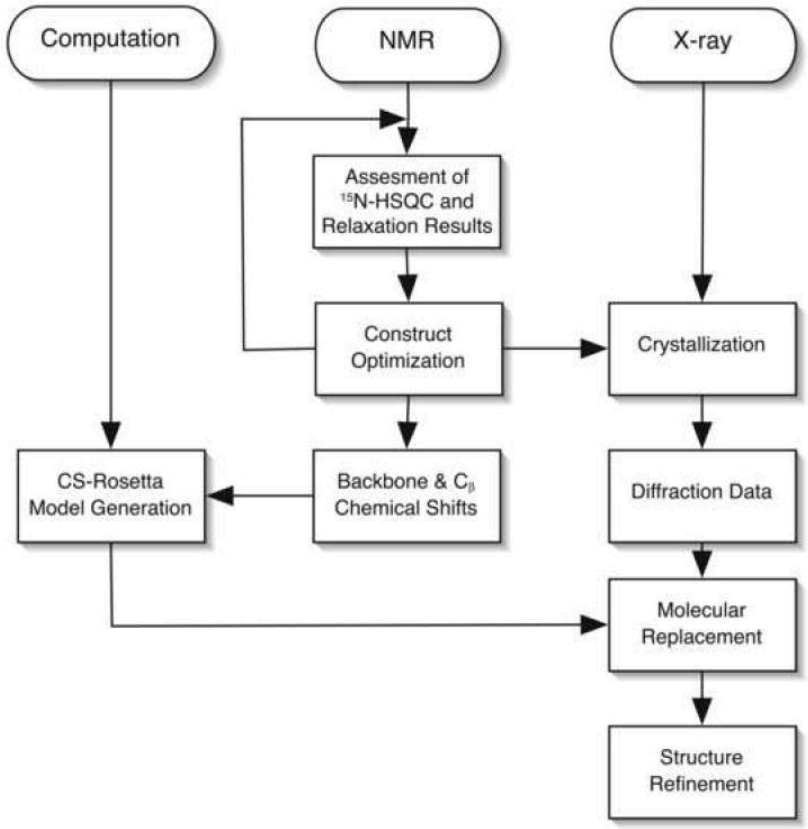
Interconnectedness:

The Human Genome Project provides the foundation for understanding the genetic blueprint, and AlphaFold leverages this blueprint to explore the functional role of proteins encoded by the genome.

Impact:

Both projects have significantly impacted scientific research and understanding of human biology, with AlphaFold building upon the insights gained from the Human Genome Project.

https://deepmind.google/discover/blog/alphafold-reveals-the-structure-of-the-protein-universe/

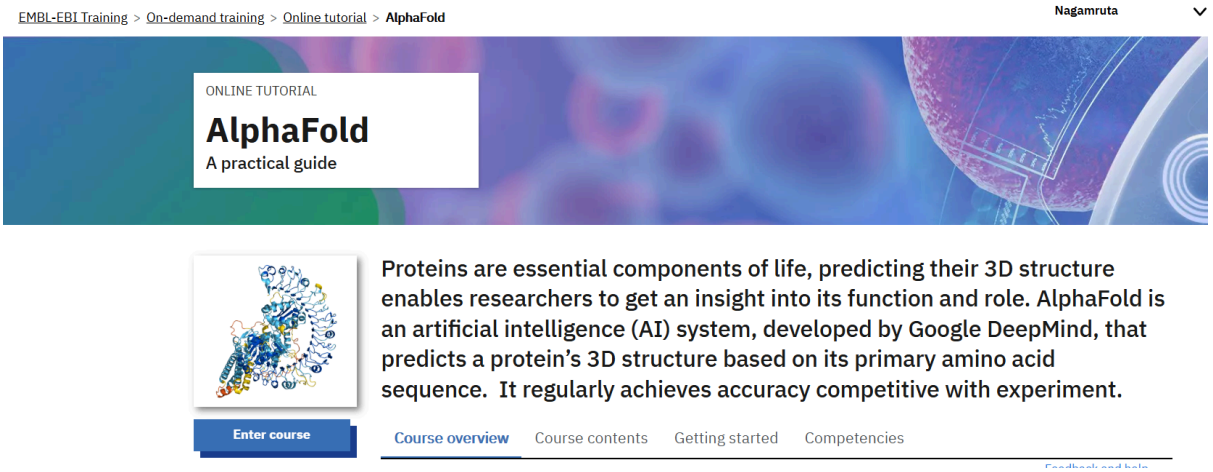


Synergetic Approach to Structure Determination. A flowchart describing the strategy for structure determination using a combination of computational techniques, NMR and X-ray Crystallography.

---

https://alphafold.ebi.ac.uk/

https://www.ebi.ac.uk/training/user/login **- Register and go through the free course**



Strengths and Limitations of AlphaFold2

✅ What AlphaFold2 Can Predict:

- **Single protein chains**
- **Protein multimers** (homo- and hetero-multimers)
- **Multisubunit protein-protein complexes**
- **Intrinsically disordered regions** (via confidence scores)
- **Novel protein folds**

⚠️ **What AlphaFold2 Struggles With:**

- **Multiple conformations** for the same sequence (predicts static structures)
- **Effects of point mutations**
- **Highly variable sequences** (e.g., antibodies)
- **Orphan proteins** (few known relatives)

❌ **What AlphaFold2 Can't Predict:**

- **Nucleic acid structures** (DNA/RNA)
- **Antigen-antibody interactions**
- **Protein-DNA and protein-RNA complexes**
- **Ligand and ion binding**
- **Post-translational modifications**
- **Membrane orientation of transmembrane proteins**
- **https://alphafold.ebi.ac.uk/faq#faq-3**

Feature	Jupyter Notebook	Google Colab
Installation	**Requires local setup (Miniconda, Anaconda, or standalone Jupyter installation)**	**No installation needed, runs in the cloud**
Accessibility	**Works offline**	**Requires an internet connection**
Collaboration	**Manual sharing (via GitHub, Google Drive, etc.)**	**Easy sharing via Google Drive**

Storage	**Uses local disk space**	**Uses Google Drive or temporary session storage**
Performance	**Limited by local machine resources**	**Offers free GPU & TPU for heavy computations**
Libraries	**Must be installed manually**	**Pre-installed ML & scientific libraries**
Security	**More control over data**	**Runs in Google's cloud, subject to their policies**
Best For	**Custom environments, offline work, sensitive data**	**Quick prototyping, machine learning, and team projects**

Create a Google Colab account

Welcome to Colab!

https://colab.research.google.com/#scrollTo=Wf5KrEb6vrkR

- **Create an API key.**

Colab now has AI features powered by Gemini.

What is Colab?

Colab, or "Colaboratory", allows you to write and execute Python in your browser, with

- Zero configuration required
- Access to GPUs free of charge
- Easy sharing

Whether you're a **student**, a **data scientist** or an **AI researcher**, Colab can make your work easier.

Step 1: Open Google Colab

1. Go to Google Colab.This notebook is a Python Jupyter notebook, and you will be running it on Google Colab.
2. Log in with your Google account.
3. Create a new notebook by clicking on **File > New Notebook**.

Once you are ready with your notebooks, we can start writing Python codes from the next session!