Exploring 3D Molecular Structures with iCn3D ProteinCodeathon@ISMB2022

Alexa M. Salsbury, Ph.D.

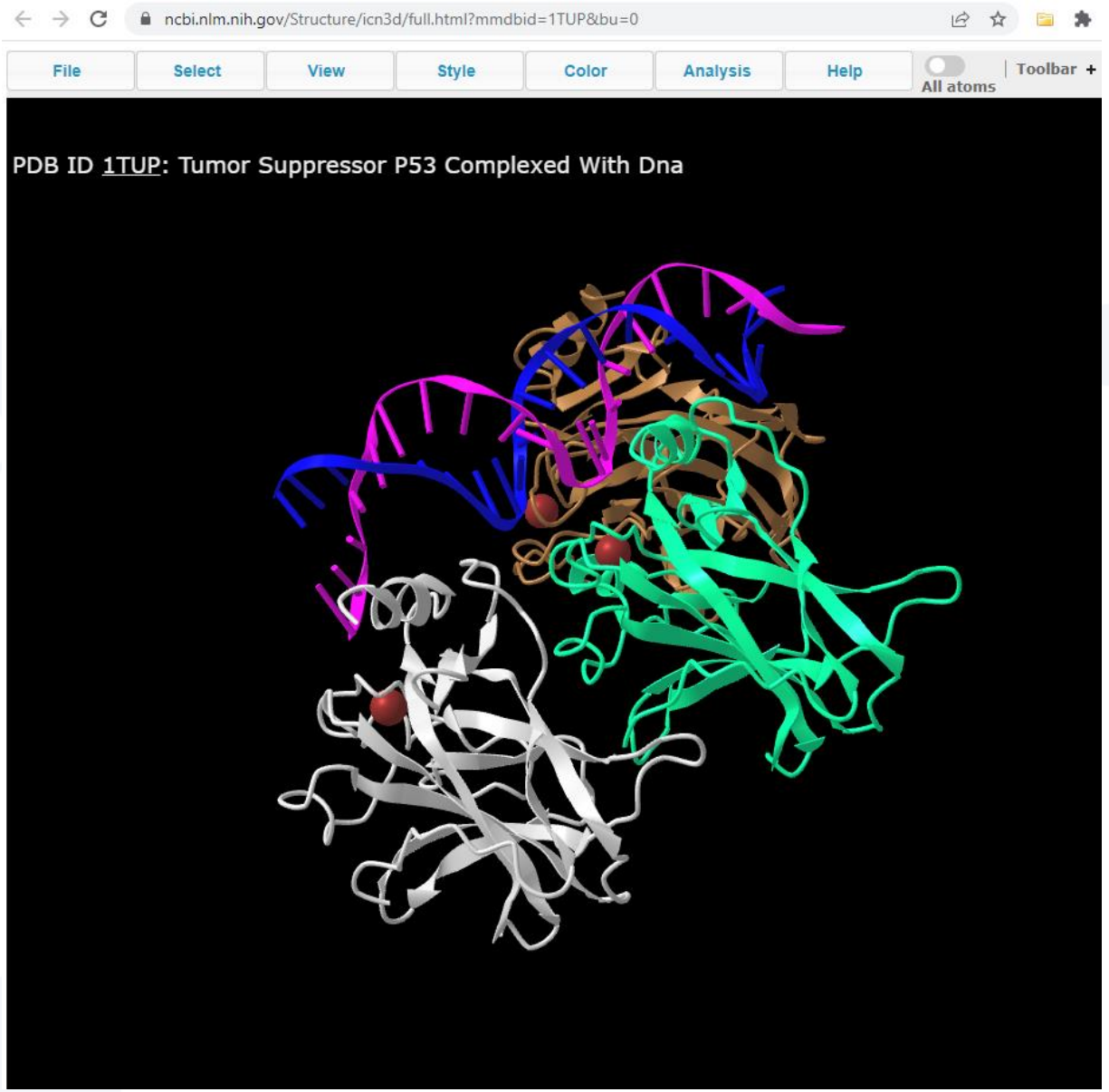
National Library of Medicine
National Center for Biotechnology Information

Overview

- Introduction
- iCn3D Orientation
  - Selection
  - Coloring
  - Style
- Interaction networks
- Mutation Analysis
- Continued learning materials

iCn3D

- Interactive, web-based 3D structure viewer
  - No installation needed!
- Users can
  - Visualize structure in 1D, 2D, and 3D
  - View sequence and structure alignments
  - Probe perturbations
  - Save/share links of their customized display

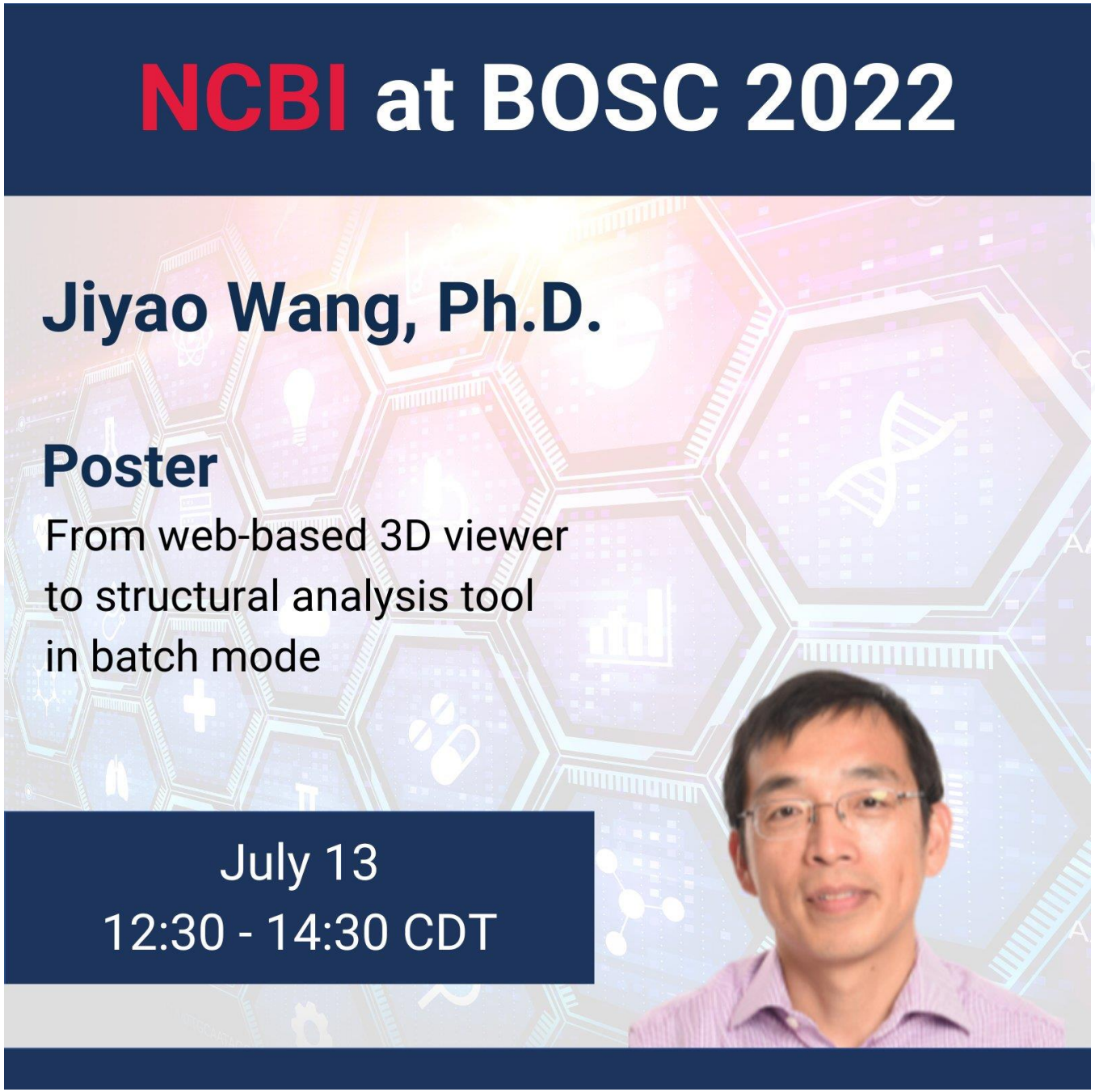


iCn3D Features of Interest

- Use iCn3D in Jupyter Notebook: pypi.org/project/icn3dpy
- 3D printing: structure.ncbi.nlm.nih.gov/icn3d/share.html?wt4TDqzhC2rhCYTD7
- Contact map: structure.ncbi.nlm.nih.gov/icn3d/share.html?rnMbe26tNsAjJLGK9
- Precalculated symmetry: structure.ncbi.nlm.nih.gov/icn3d/share.html?bGH1BfLsiGFhhTDn8
- Symmetry dynamically: structure.ncbi.nlm.nih.gov/icn3d/share.html?6NvhQ45XrnbuXyGe6
- Electron density map: structure.ncbi.nlm.nih.gov/icn3d/share.html?QpqNZ3k65ToYFvUB6
- EM map: structure.ncbi.nlm.nih.gov/icn3d/share.html?L4C4WYE85tYRiFeK7
- Transmembrane protein: structure.ncbi.nlm.nih.gov/icn3d/share.html?jMN16mJyR9STUx6E6
- Solvent Accessible Area: structure.ncbi.nlm.nih.gov/icn3d/share.html?xKSyfd1umbKstGh29

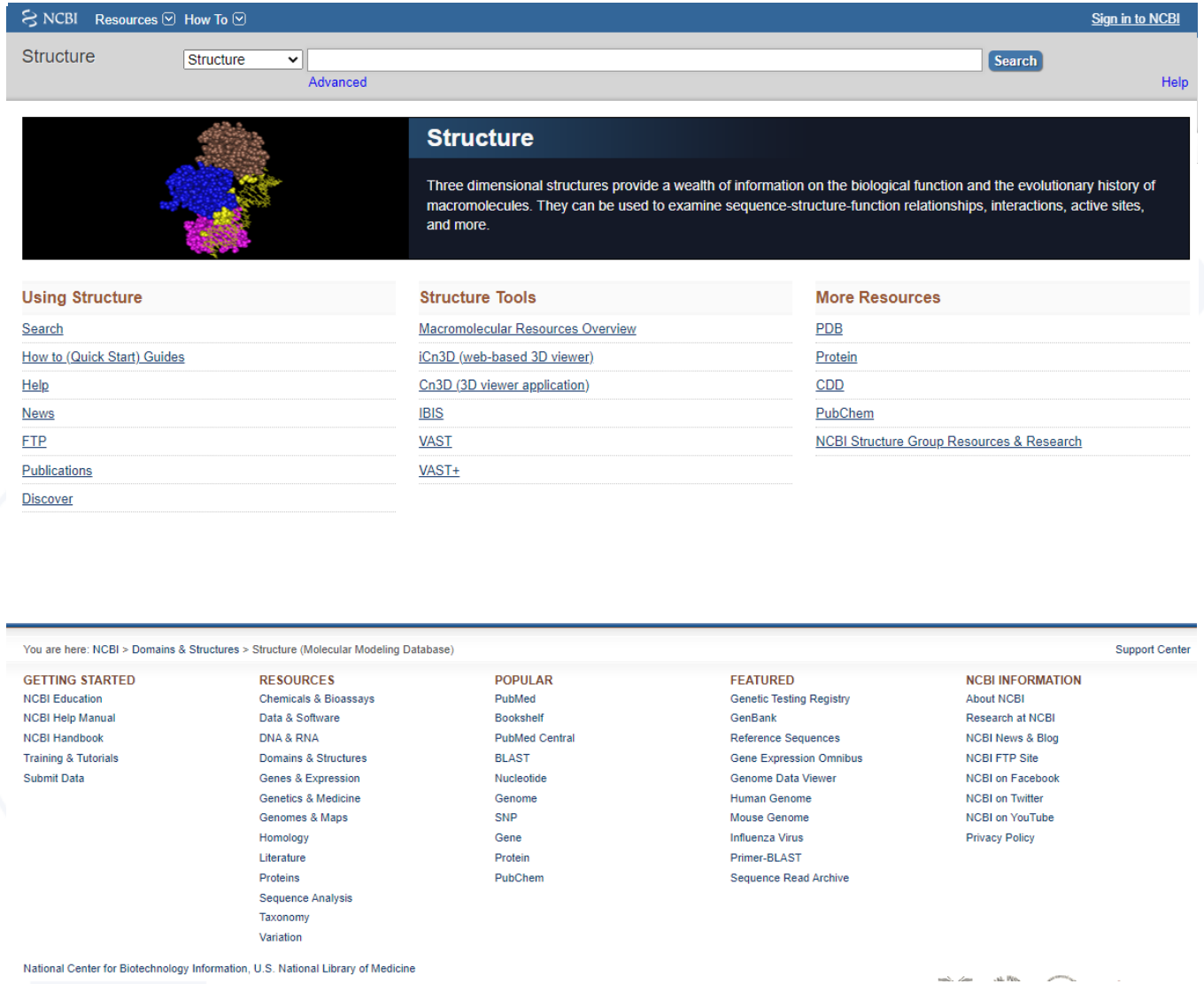
New Features!

- Virtual reality view
- Batch analysis



Structure Database

- Updated monthly
- Derived from PDB records
- Additional information added, including:
  - Explicit chemical graph information
  - Validation (secondary structure elements)
  - Includes taxonomy
- Connects 3D to associated literature, molecular data, chemical data, and other NCBI tools



iCn3D Shortcuts

Rotate

- **Left mouse button** can be used to rotate the structure
- **Key L** - left
- **Key J** - right
- **Key I** - up
- **Key M** - down

Zoom

- **Middle mouse button** can be used to zoom
- **Left Mouse + Shift** can be used as an alternative to the middle mouse button
- **Key Z** - zoom in
- **Key X** - zoom out

Translate

- **Right mouse button** - can be used to translate (slide) the structure to a different location within the 3D window
- **Left Mouse + Ctrl** can be used as an alternative to the right mouse button
- **Arrow Left** - left
- **Arrow Right** - right
- **Arrow Up** - up
- **Arrow Down** - down

Select

- **Alt + Click (PC)** or **Option + Click (Mac)**- can be used to select atom/residue/strand , hold **Ctrl + Click** to add another

iCn3D Demo (1/5)

- Click **`full-feature 3D viewer`** on the Molecular Graphic
- Or go to https://www.ncbi.nlm.nih.gov/Structure/icn3d/full.html
  - Input **`1TUP > Load Biological Unit`**
- Orient yourself (see iCn3D Shortcuts for help)
- Hover over structure with your mouse to view residues
- **`Select > Select on 3D > Atom`** to see atomistic details
- You can revert to selecting by residues **`Select > Select on 3D > Residue`**

https://structure.ncbi.nlm.nih.gov/icn3d/share.html?PrJm9U7w2irwV2pT7

iCn3D Demo (2/5)

- Change styling with **`Style > Sidechains > Lines`**
- Explore different **`Style`** options
- Explore different **`Color`** options
  - **`Chain`** – default, colors structural components differently
  - **`Rainbow`** – N-term or 5’ end is red and flows to blue for C-term or 3’ end
  - **`Charge`** – colors positively charged as blue, negatively charged as red, and neutral as gray
  - **`Atom`** – colors C gray, O red, N blue, S yellow
  - **`Secondary`**, **`Hydrophobicity`**, and **`Solvent Accessibility`** options are useful for more in-depth analysis of structure

https://structure.ncbi.nlm.nih.gov/icn3d/share.html?GkfzpCjxnCF8L79o6

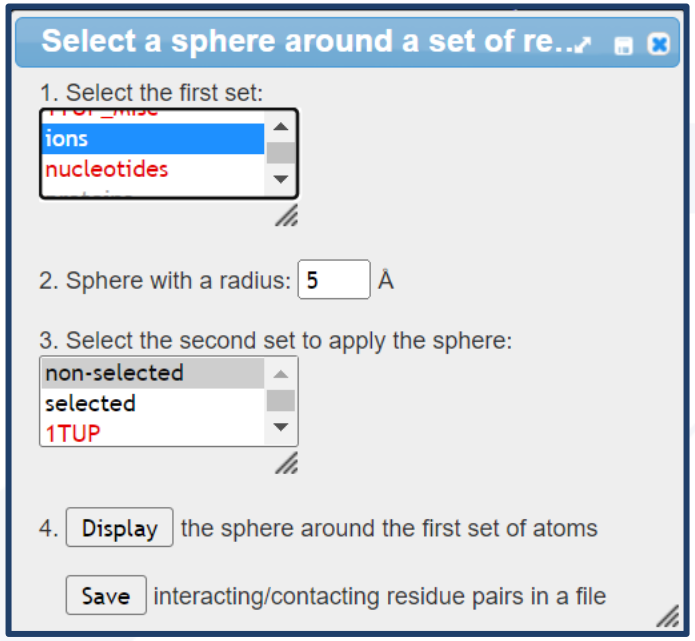
iCn3D Demo (3/5)

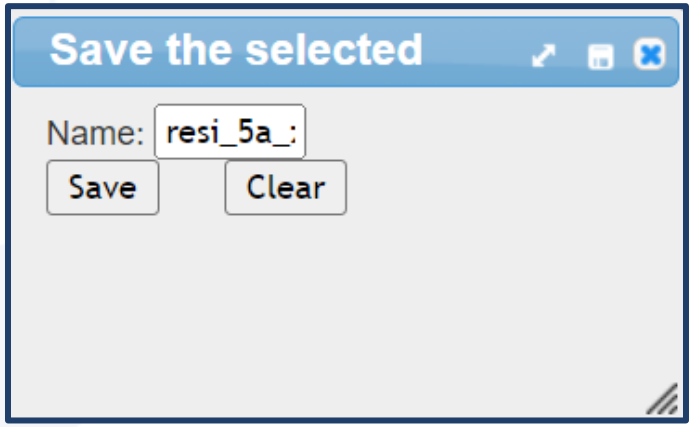
- Make specific selections with **`Analysis > Defined Sets`**
- Select **`nucleotides`** and change color with **`Color > Rainbow`**
- Select residues by sequence with **`Analysis > Seq. & Annotations`**
  - Uncheck annotations and click **`Details`**
  - Highlight residues from the sequence to select
- Close or minimize **`Seq. & Annotations`** when not using

https://structure.ncbi.nlm.nih.gov/icn3d/share.html?F9ZUDzWifapM8ZiXA

iCn3D Demo (4/5)

- Explore ion interactions with **`Select > By Distance`**
  - Choose **`ions`**
  - Set sphere radius to **`5 Å`**
  - Choose **`non-selected`**
- Change the style of these residues with **`Style > Proteins > Stick`**
- Save the selection by **`Select > Save Selection`** and give name like **`resi_5a_zn`**





https://structure.ncbi.nlm.nih.gov/icn3d/share.html?qmYR6yC46RHsUYMe8

iCn3D Demo (5/5)

- View the ion interactions `Analysis > Defined Sets > ions` and your newly named selection `resi_5a_zn` and `View > View Selection`
- Change your background color with `Style > Background`
- Saving your files
  - As a PNG with `File > Save Files > iCn3D PNG image`
  - As an interactive link with `File > Share Link` and copy
  - For 3D printing with `File > 3D Printing`

https://structure.ncbi.nlm.nih.gov/icn3d/share.html?PwwbXdh9fnBs6gyy7

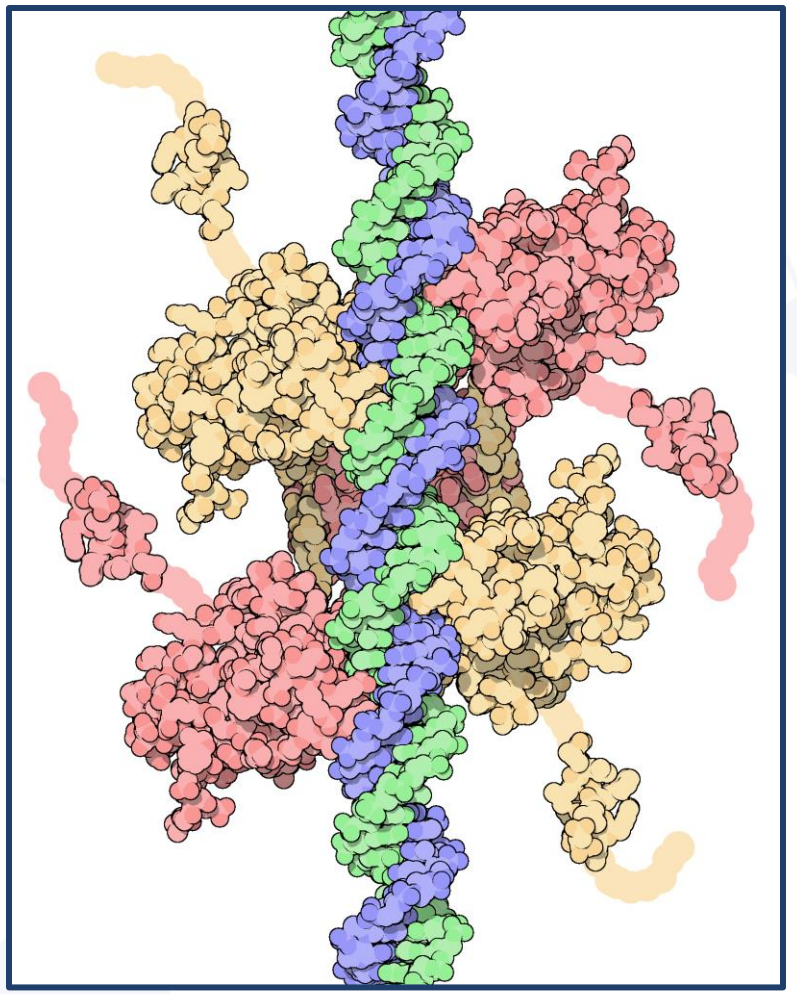
iCn3D Exploration

5-minute exercise!

- If you select something accidentally `Select > Clear Selection`
- If you need to undo `View > Undo`
- Get additional help by:
  - `Show Help > Help Docs`
  - `Help > Selection Hints`
- Like what you've rendered? Share your interactive link in the chat!
- Ask us!

P53 DNA-binding

- P53 binds to regulatory sites in the genome and:
  - Initiates protein production that stops cell division until damage is repaired
  - Initiates apoptosis
- Rich in + charged amino acids (Arg, His, Lys)
  - + charged amino acids commonly interact with negatively charged nucleic acid backbones
  - Usually interact at the major grove



P53 DNA- binding (from PDB)

Mutation Example 1

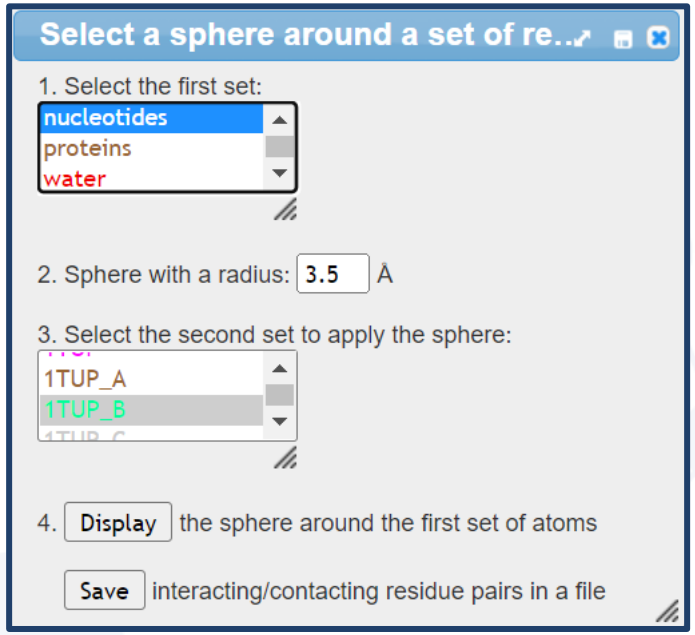
- Go to back to iCn3D
- What residues seem important to P53-binding domain?

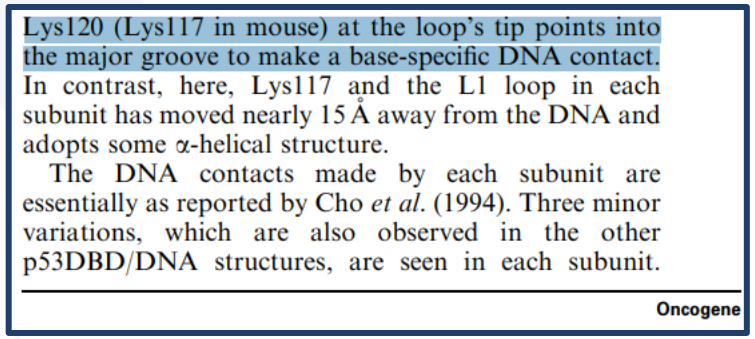
  `Select > by Distance`

- Take a closer look at these residues

  `Style > Protein > Stick`

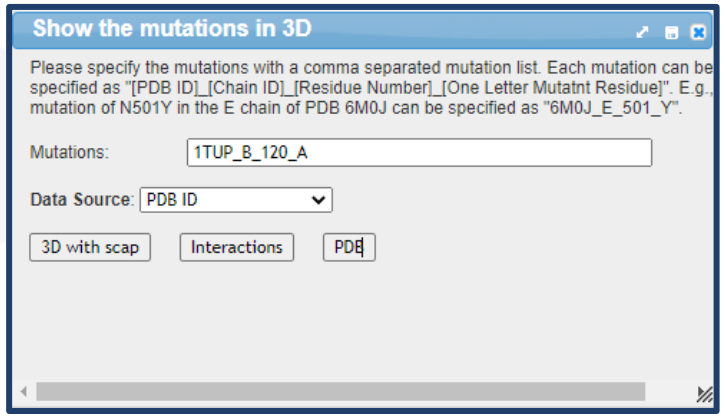
- Lys120 points into the major groove to make bas-specific contact. How might a mutation affect interactions?
  - `Analysis > Mutation > 1TUP_B_120_A` and select `Interactions`




Lys120 (Lys117 in mouse) at the loop's tip points into the major groove to make a base-specific DNA contact. In contrast, here, Lys117 and the L1 loop in each subunit has moved nearly 15 Å away from the DNA and adopts some α-helical structure.

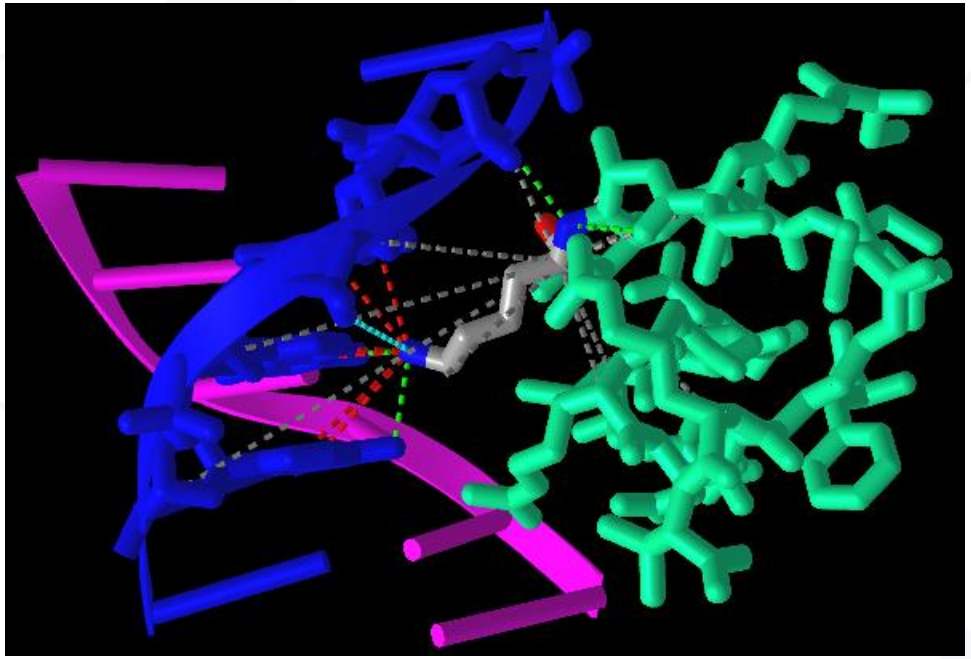
The DNA contacts made by each subunit are essentially as reported by Cho *et al.* (1994). Three minor variations, which are also observed in the other p53DBD/DNA structures, are seen in each subunit.

Oncogene

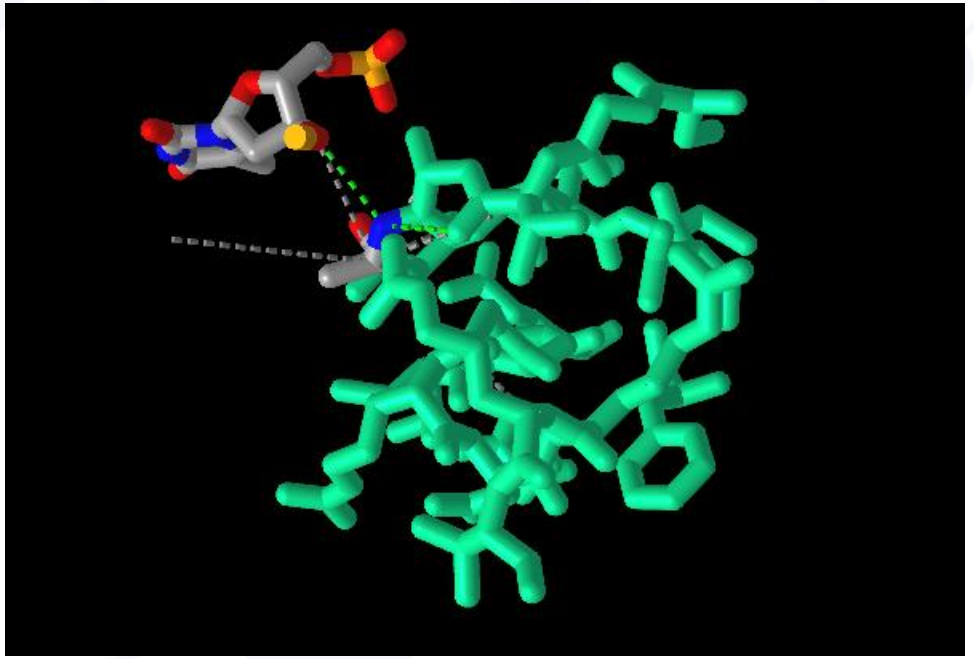


Mutation Discussion 1

- K120A mutation results in loss of interaction with the DNA groove



Wild type K120



Mutant A120

Mutation Example 2

5-minute exercise!

- Literature shows that Arg248 and Arg273 are common P53 mutations implicated in disease
- Use the Mutation analysis to understand how these mutations may affect interactions

```
1TUP_B_248_W
1TUP_B_273_H
```

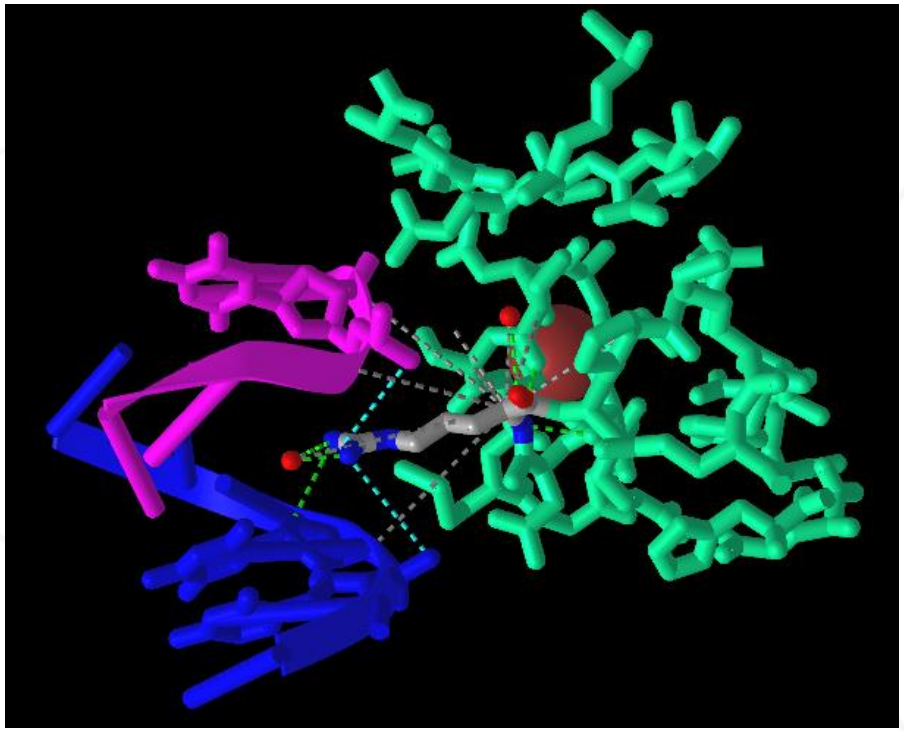
Distinct pattern of p53 phosphorylation in human tumors

Phosphorylation of mutant p53 in tumor-derived cell cultures

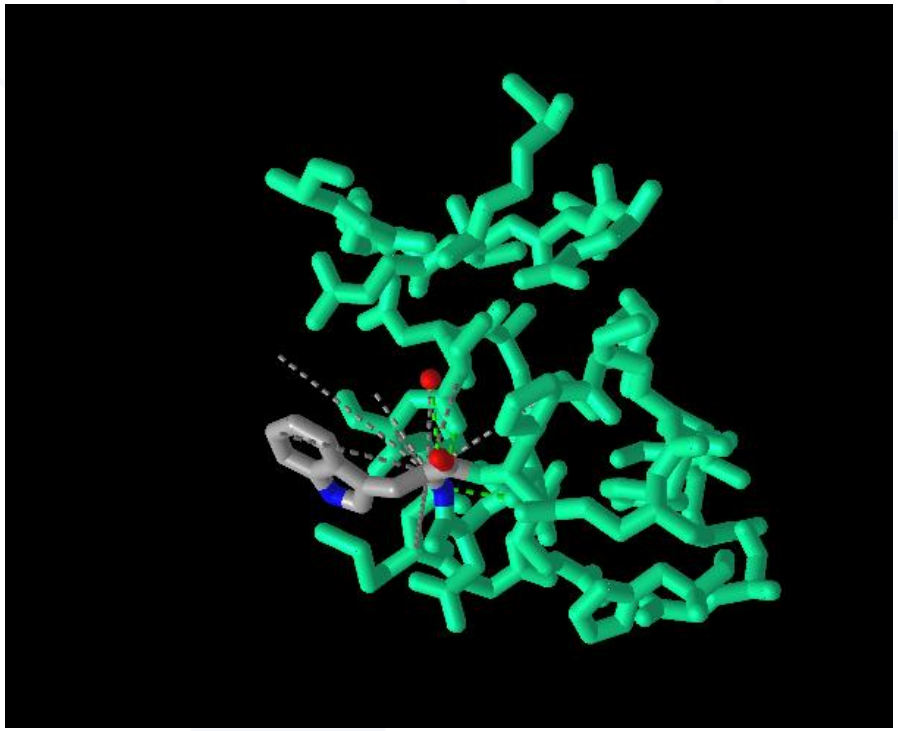
To determine the pattern of mutant p53 phosphorylation and acetylation in tumor-derived cell lines under normal growth conditions we analysed 18 cell lines with defined p53 mutations. Cell lines used for phospho-analysis were derived from seven tumor types and included a total of nine different mutations. Several tumor-derived cell lines with the same hot spot mutation (R248W or R273H) were included in this analysis to enable comparison of the phosphorylation pattern among different tumors that have the same mutation. As controls for this analysis, we used two non-transformed fibroblast cell lines (GM00038 and TIG) known to harbor wild type p53. Additional analysis has been carried out in parallel on tumor-derived tissues and cell lines that harbor wild type p53, thus allowing comparison of the phosphorylation pattern of wild type and mutant forms of p53 within the tumor environment.

Mutation Discussion 2

- R248W mutation results in loss of hydrogen and salt bridge interactions



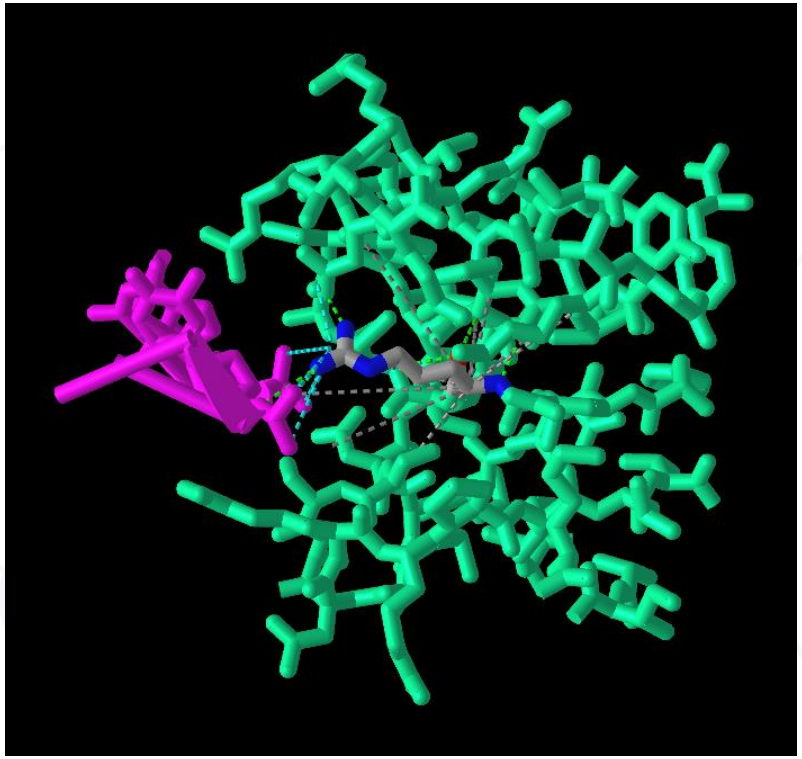
Wild type R248



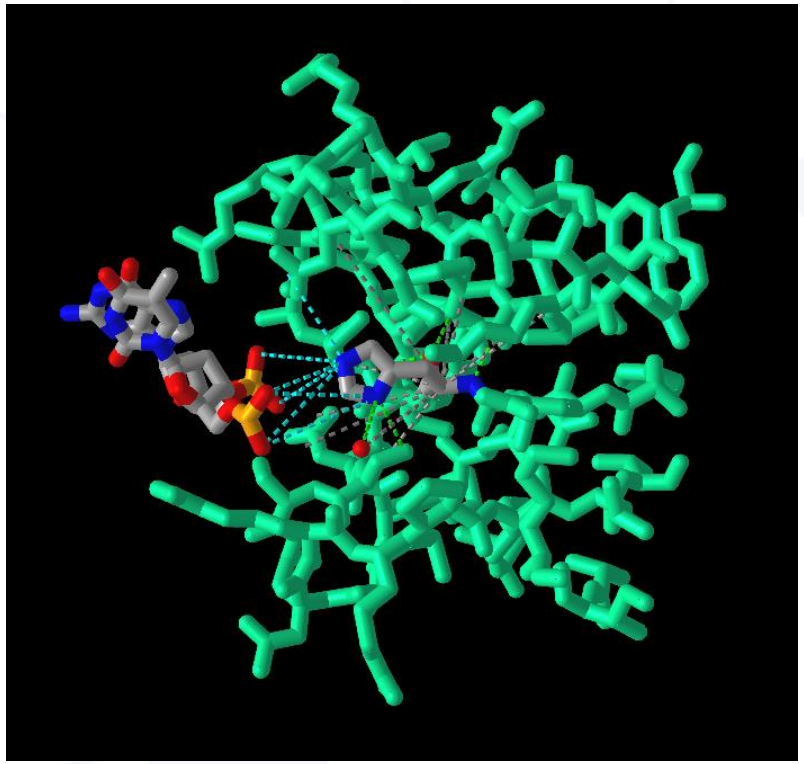
Mutant W248

Mutation Discussion 3

- R273H mutation results in loss of hydrogen bond interactions



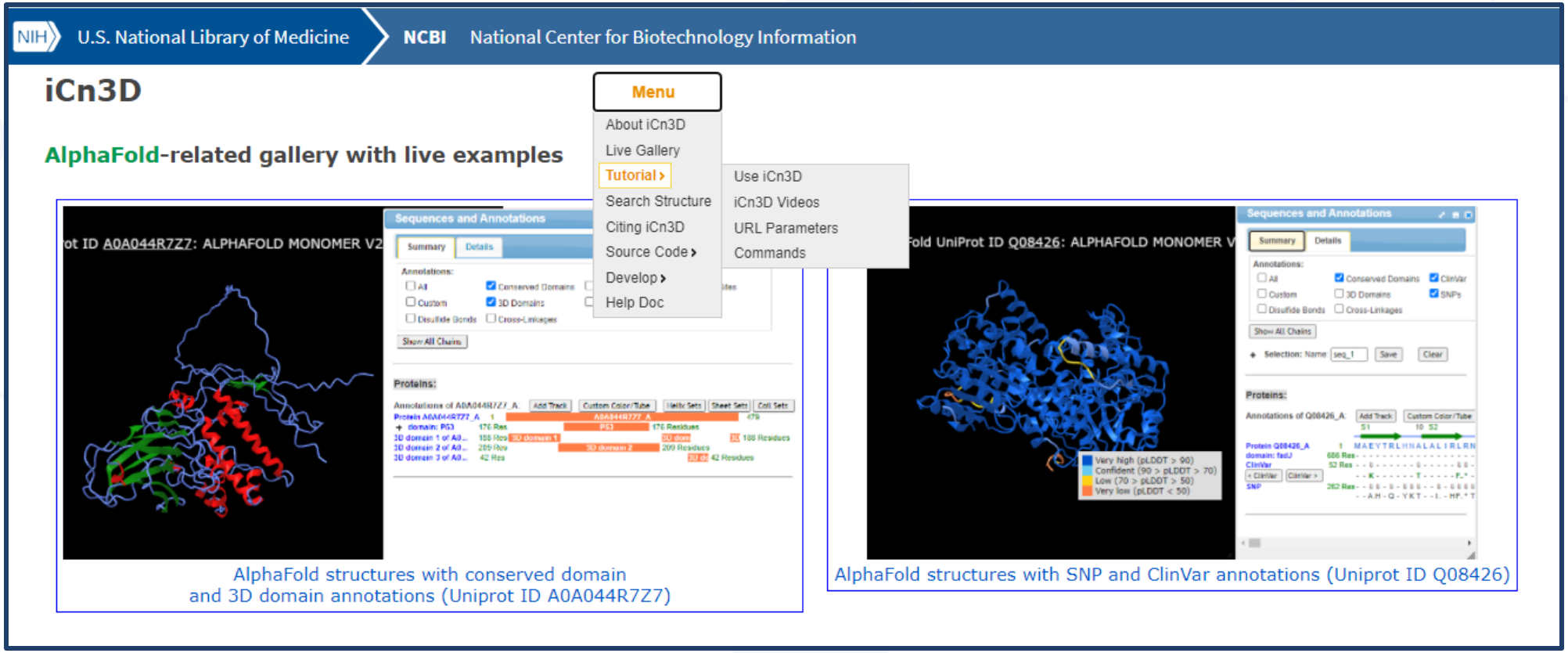
Wild type R273



Mutant H273

Continue learning about iCn3D

Tutorials and help documents are available here.



AlphaFold structures with conserved domain and 3D domain annotations (Uniprot ID A0A044R7Z7)

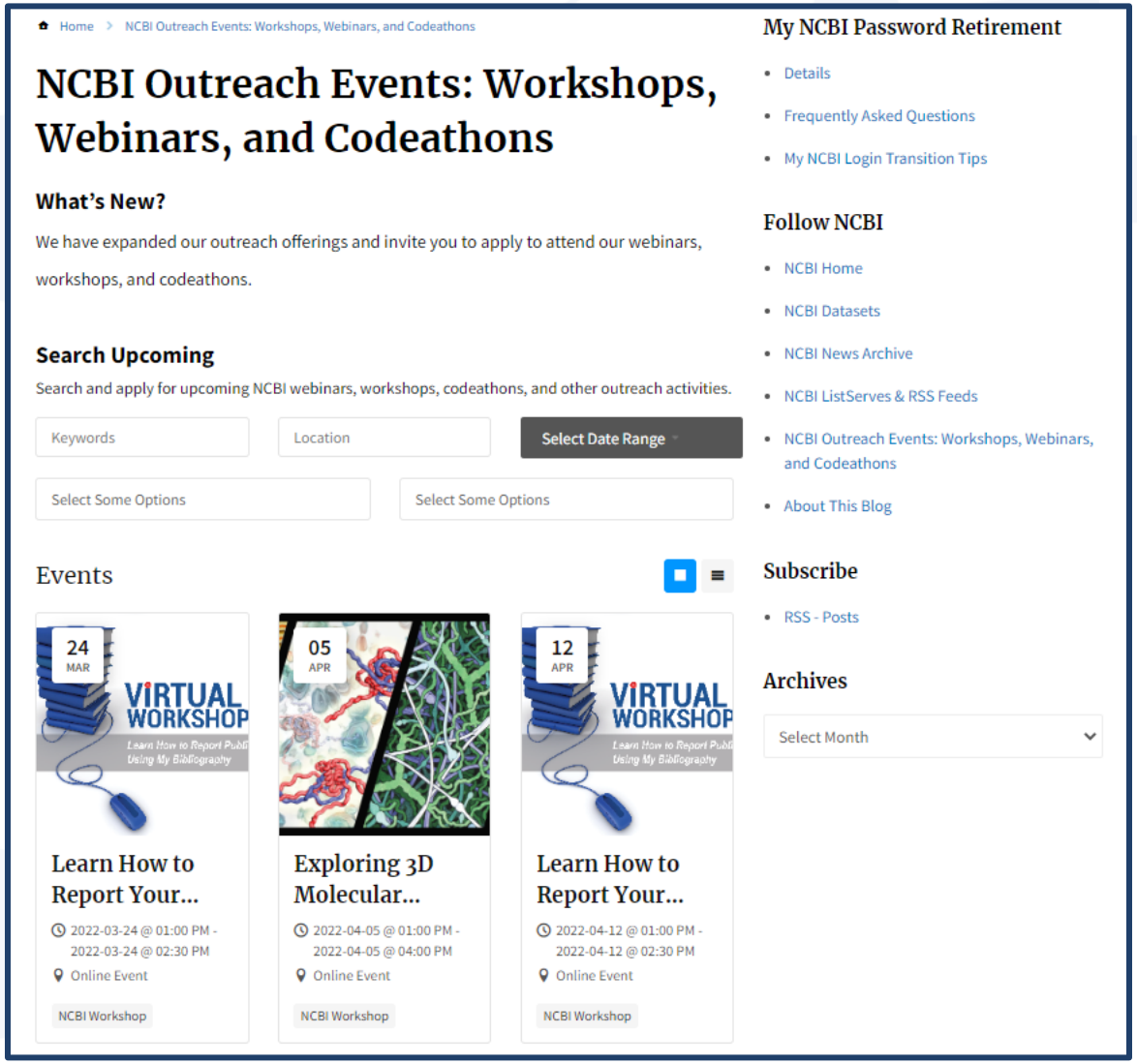
AlphaFold structures with SNP and ClinVar annotations (Uniprot ID Q08426)

Continue learning about NCBI Resources

- Join us for workshops, webinars, or codeathons!

  NCBI Insights Blog

- Follow us on social media:
  - Twitter
  - LinkedIn
  - Facebook



Exploring 3D Molecular Structures with iCn3D Supplemental Learning Materials

Alexa M. Salsbury, Ph.D.

National Library of Medicine
National Center for Biotechnology Information

Structural Biology

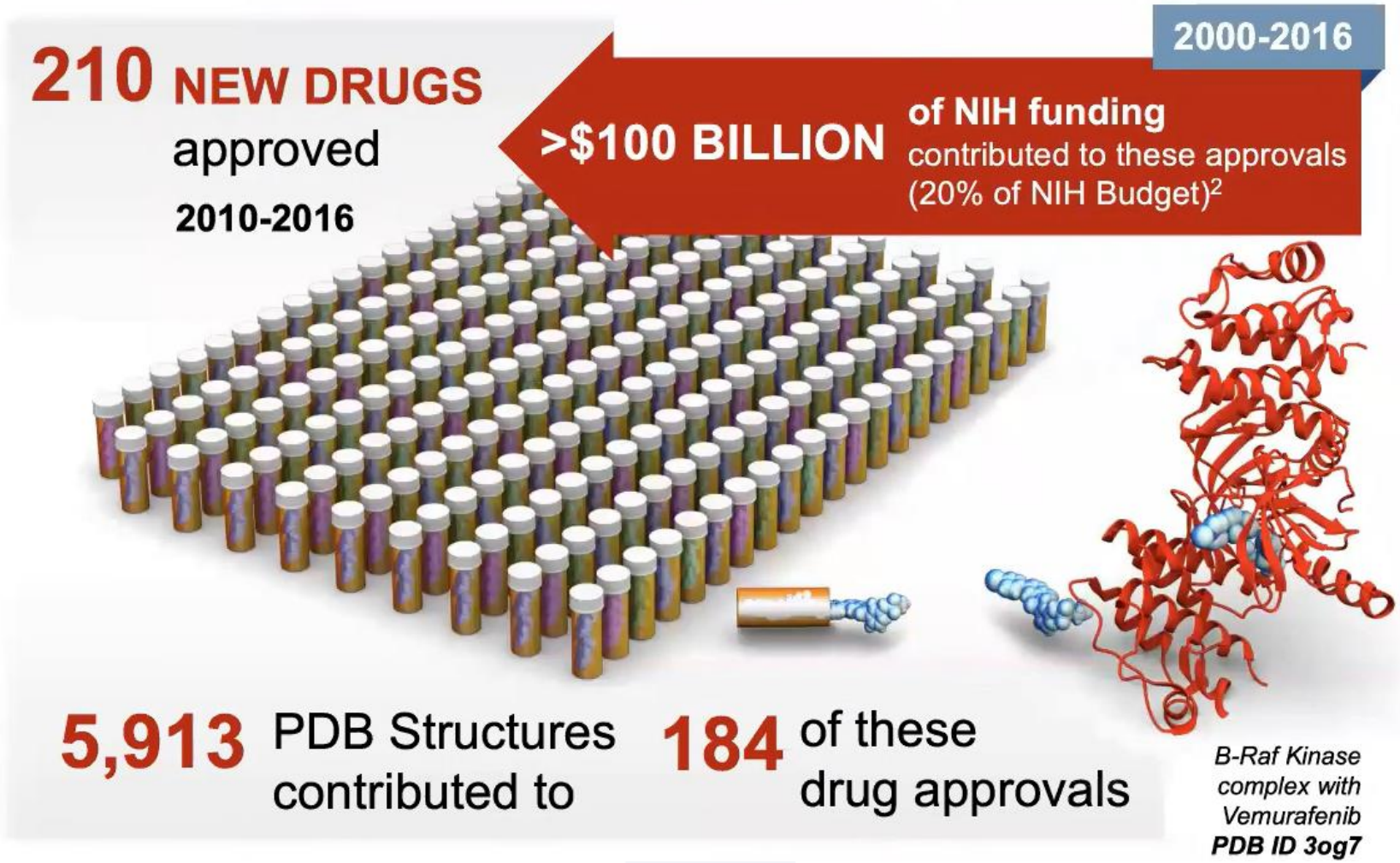
1952-1953- Pioneering DNA structure work by Wilkins, Franklin, Watson, & Crick.

1956-1960- Rich & Davies' structural experiments showed how information could be transferred from DNA to RNA.

1957- The first protein with a crystal structure was solved in by Kendrew and co-workers

Now- over 175,000 structures are publicly available!

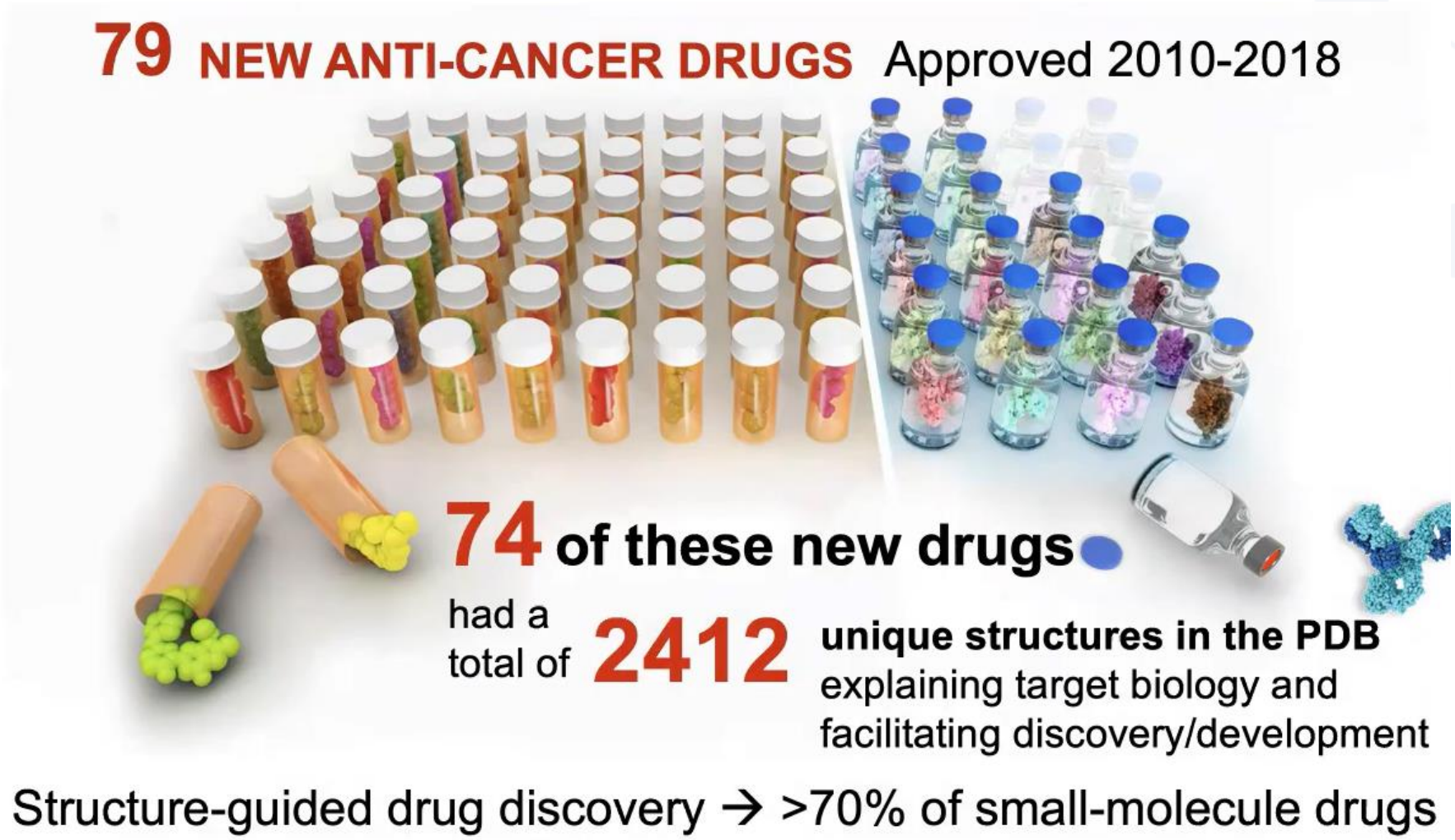
Impact of Structures on New Drug Approvals



National Library of Medicine
National Center for Biotechnology Information

Westbrook & Burley (2019) *Structure 27,* 211-217
Galkina Cleary *et al.* (2018) *PNAS 115*, 2329-2334

Impact of Structures on Anti-Cancer Drug Approvals

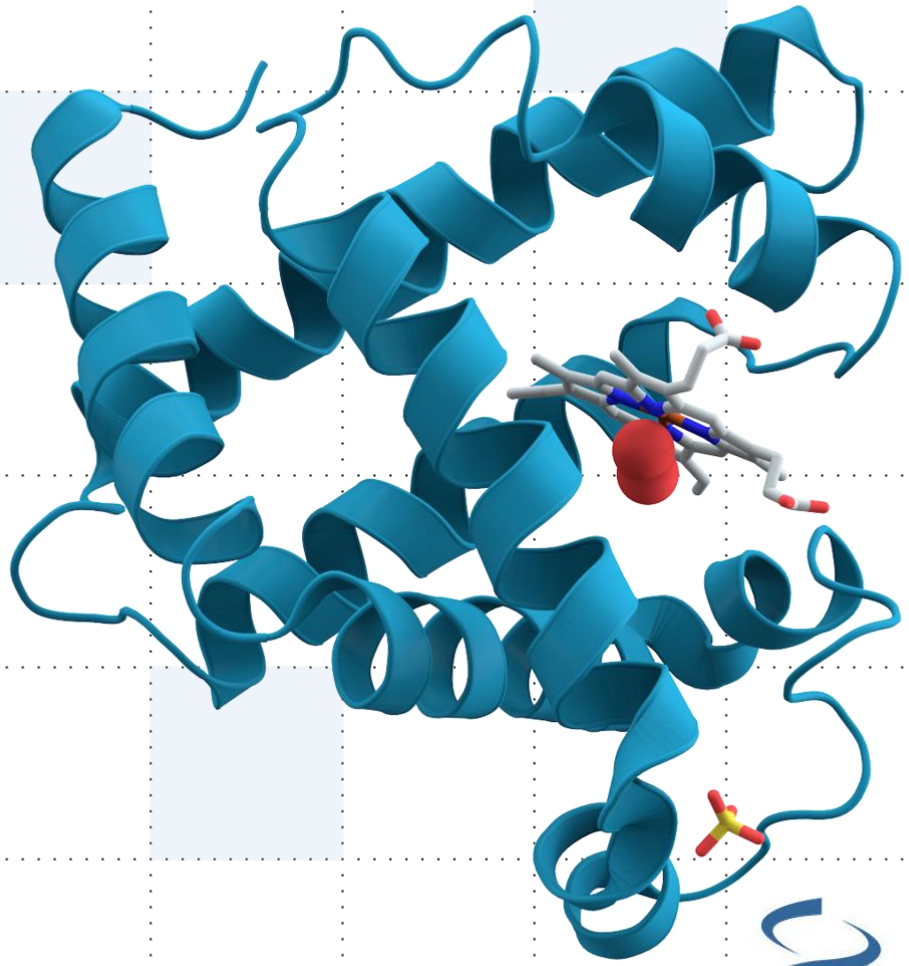


National Library of Medicine
National Center for Biotechnology Information

Westbrook & Burley (2019) *Structure 27,* 211-217

To keep you alive, your body must:

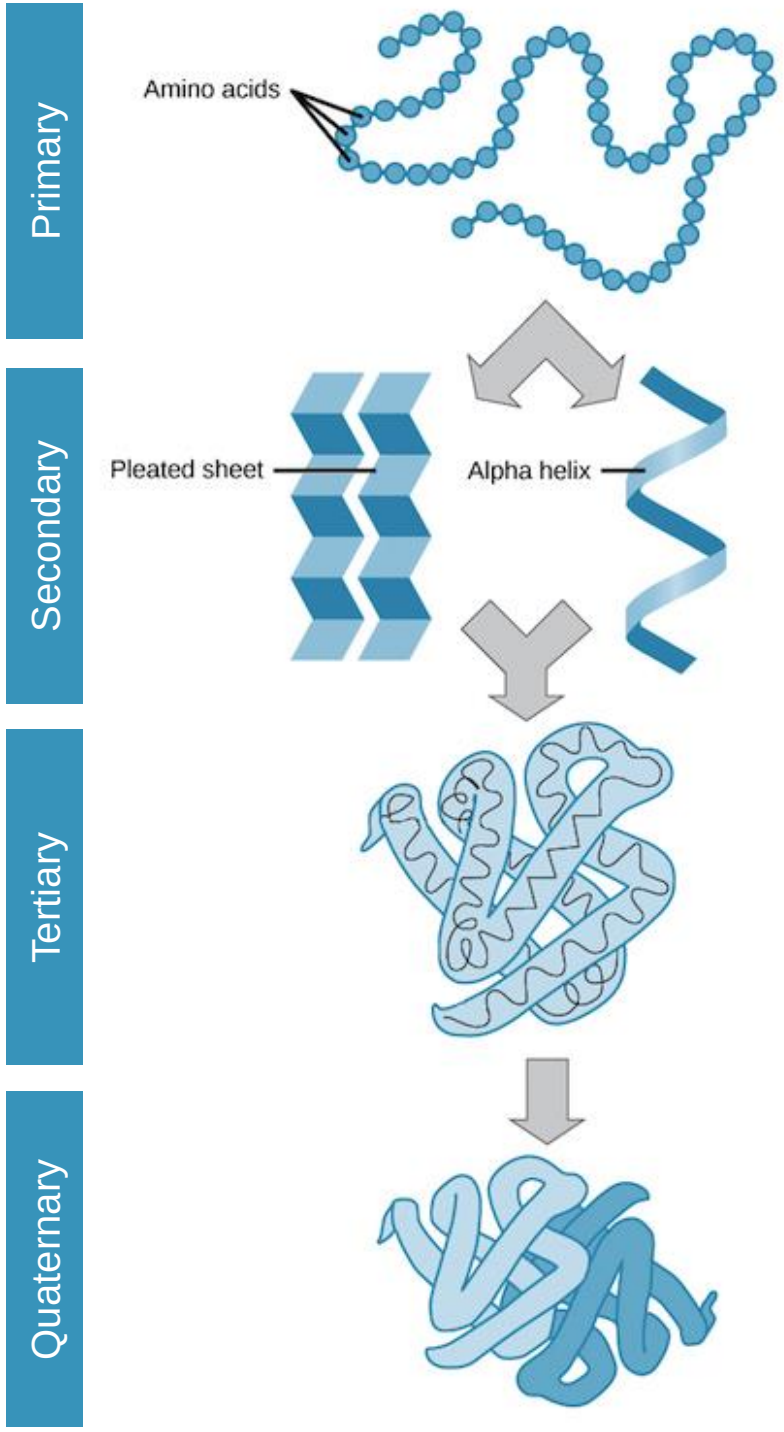
- Convert oxygen to usable energy
- Digest food and recover nutrients
- Build and maintain body tissue
- Fight off invasions and detoxify poisons



Proteins do all these things!

Protein Structure

- **Primary-** sequence of amino acids
- **Secondary-** hydrogen bonding of the peptide backbone that causes amino acids to fold into a repeating pattern
- **Tertiary-** 3D folding pattern of a protein due to side chain interactions
- **Quaternary-** protein consisting of more than one polypeptide

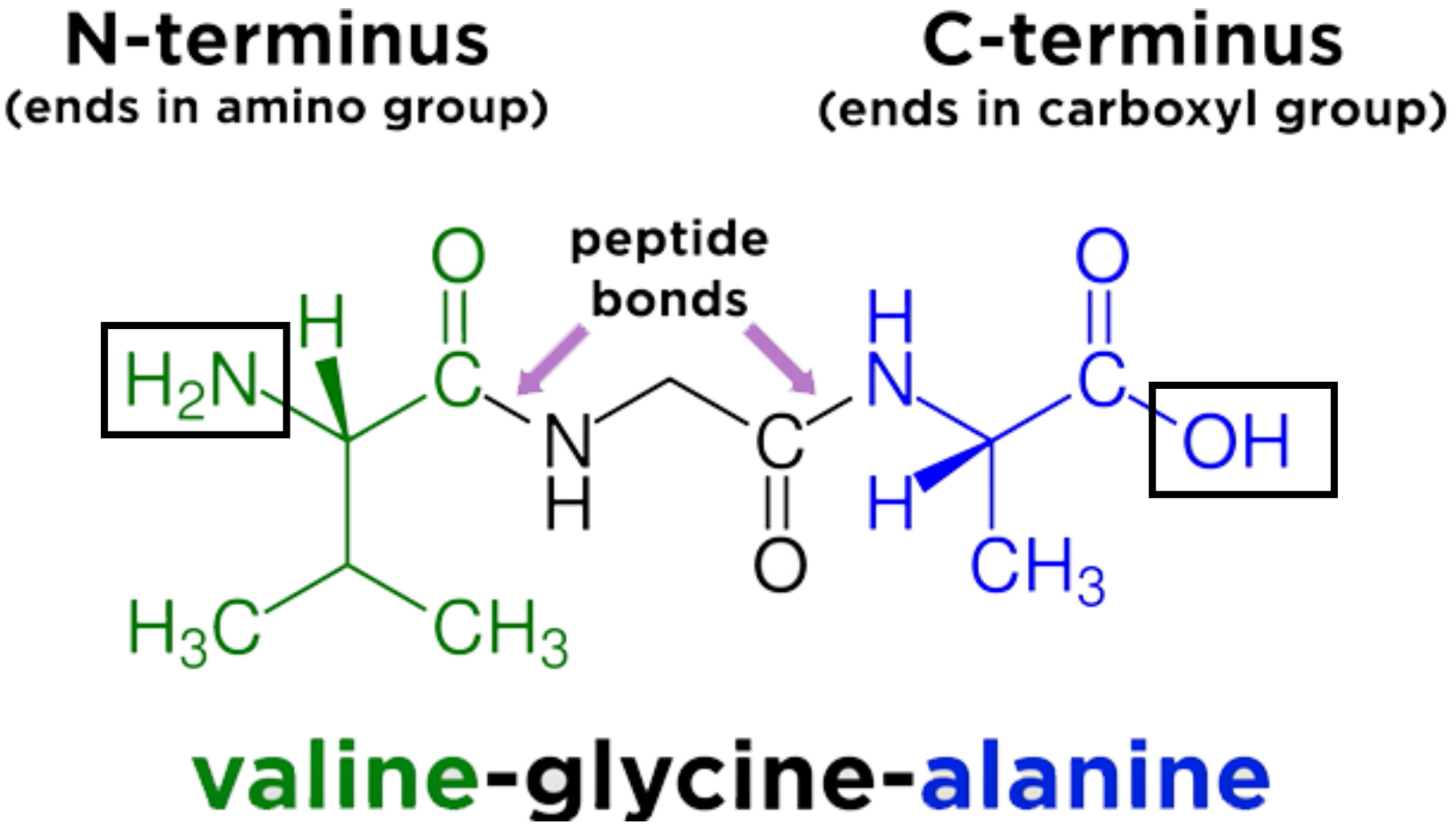


Classification of Proteins – Size

- Most peptides and proteins have 2-2000 amino acids
- Assuming average MW per amino acid of 110 daltons, MW range for peptides and proteins in range 220-220,000 daltons
- Some proteins consist of single polypeptide chain and are monomeric
- Others consist of multiple polypeptide chains (sometimes identical, sometimes not) and are oligomeric

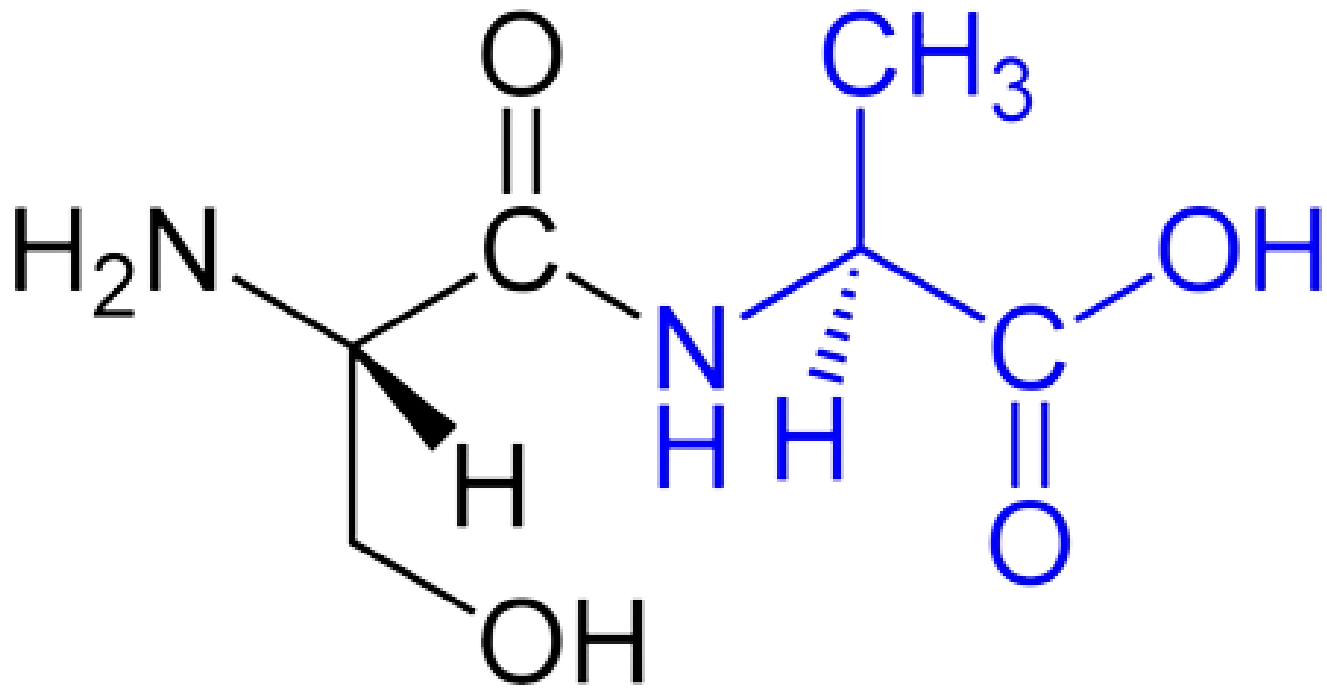
NCBI

Terminology



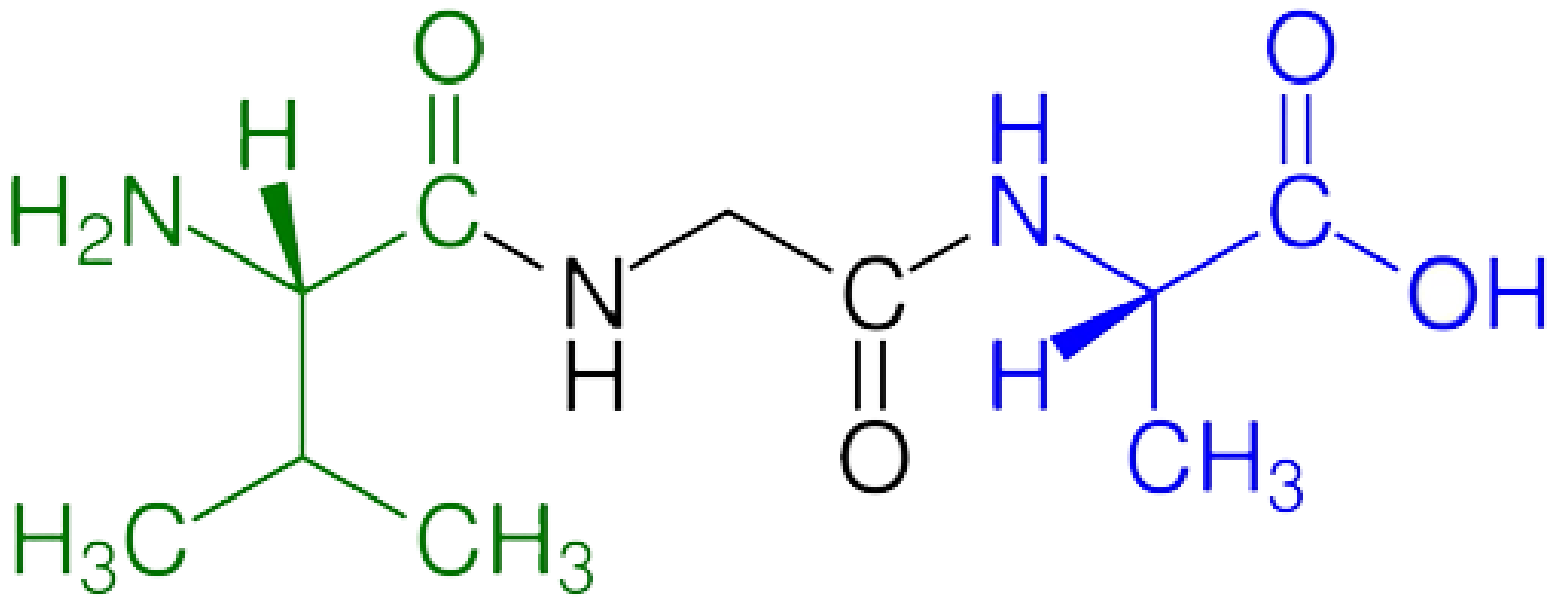
Terminology

dipeptide (2 amino acids)

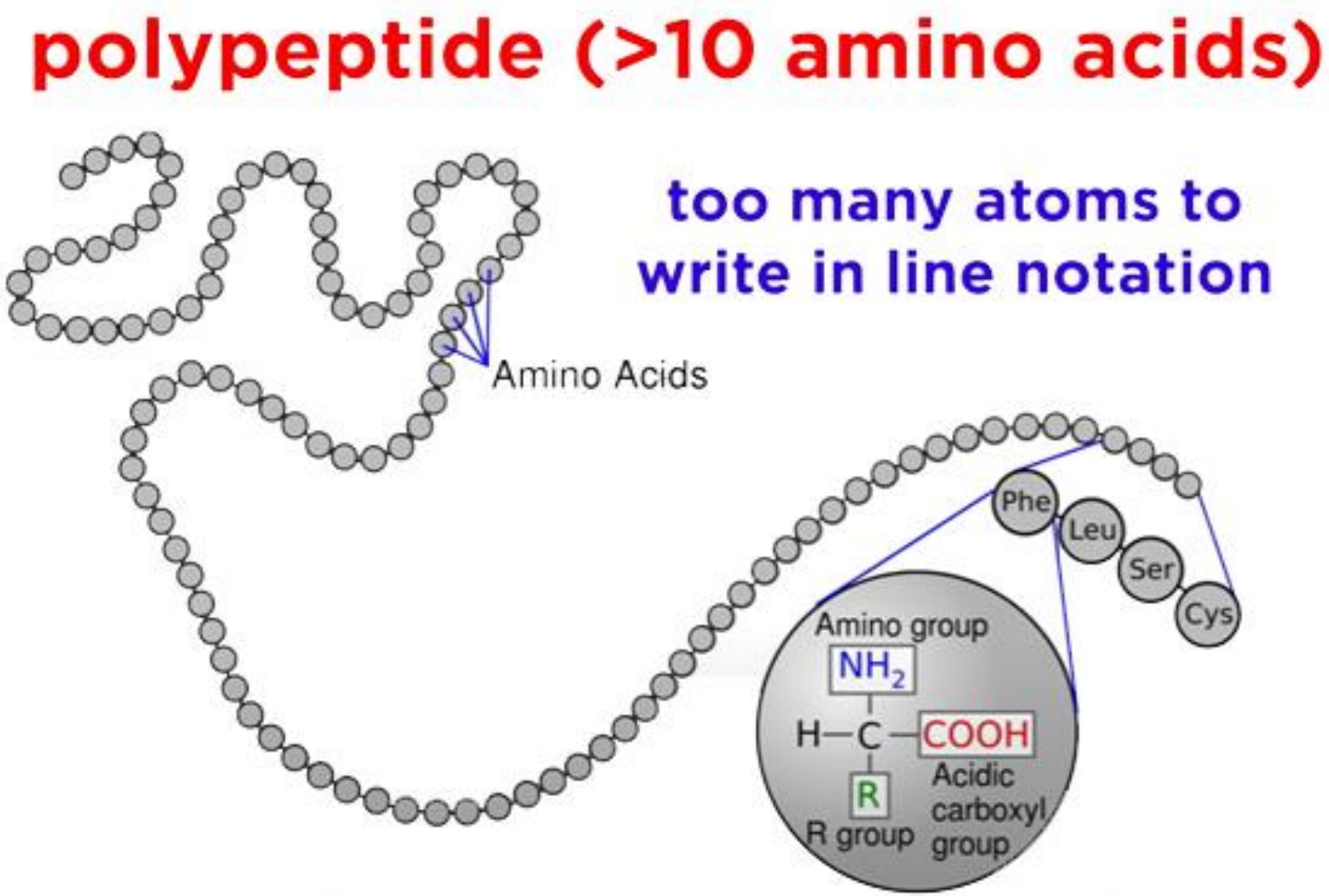


Terminology

oligopeptide (3-10 amino acids)

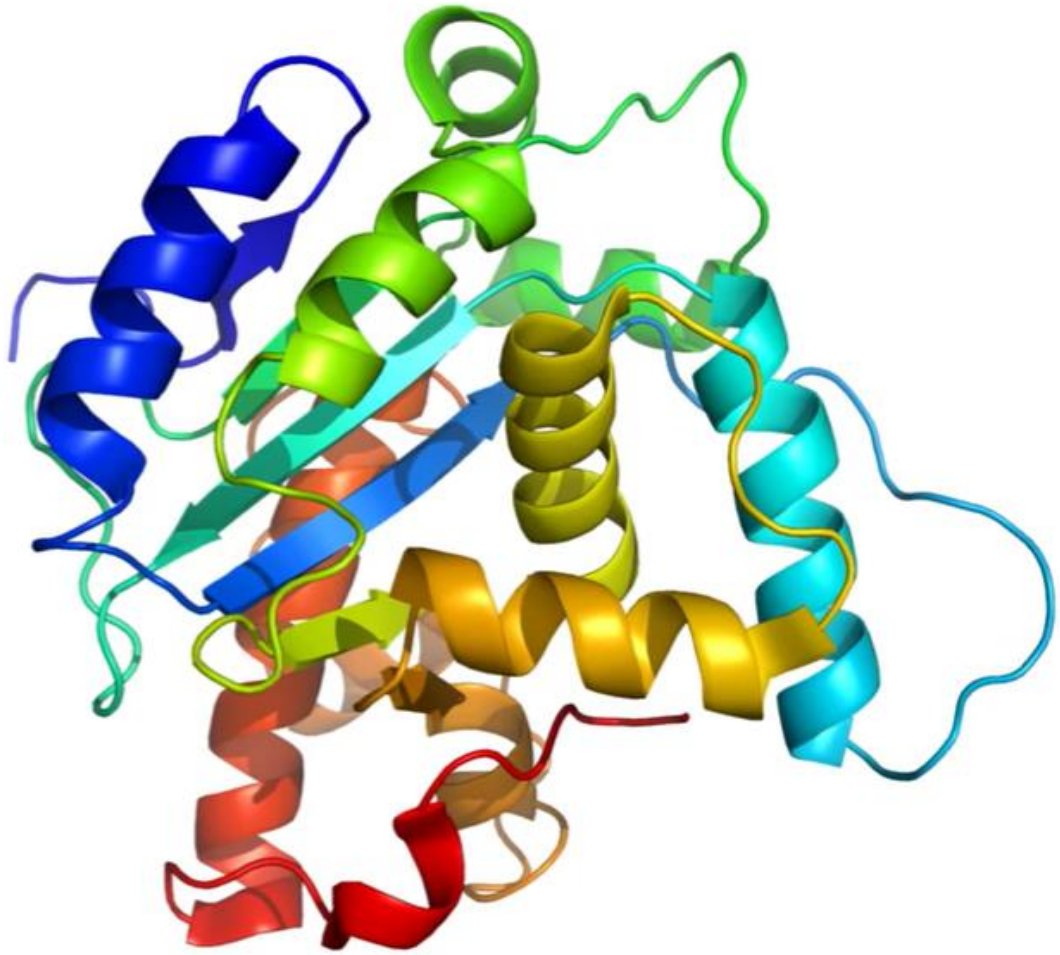


Terminology

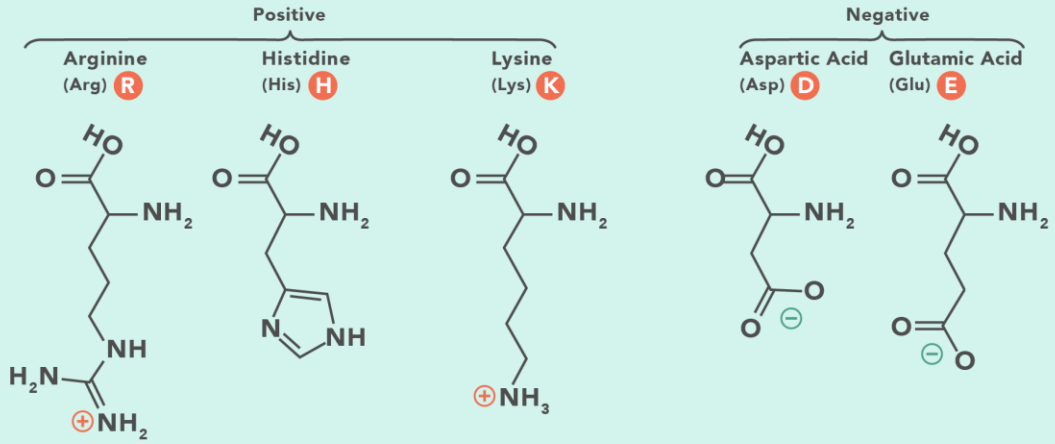


Terminology

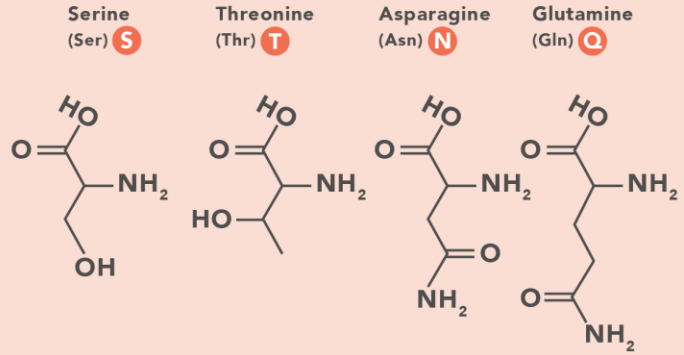
protein (generally 300-1000 amino acids)



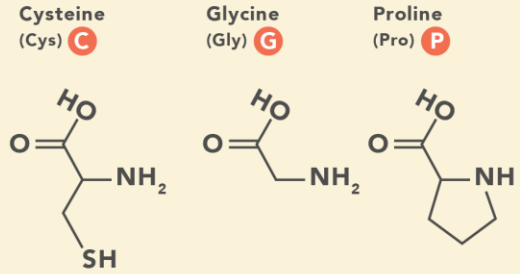
A. Amino Acids with Electrically Charged Side Chains



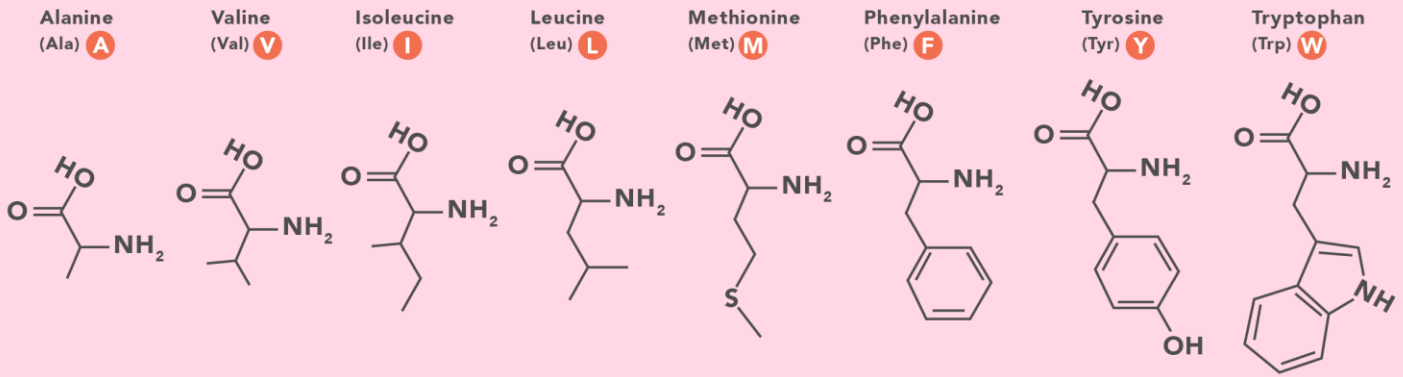
B. Amino Acids with Polar Uncharged Side Chains



C. Special Cases



D. Amino Acids with Hydrophobic Side Chains



Functional definition:

Enzymes: Accelerate biochemical reactions

Structural: Form biological structures

Transport: Carry biochemically important substances

Defense: Protect the body from foreign invaders

Structural definition:

Globular: Complex folds, irregularly shaped tertiary structures

Fibrous: Extended, simple folds -- generally structural proteins

Cellular localization definition:

Membrane: In direct physical contact with a membrane; generally water insoluble.

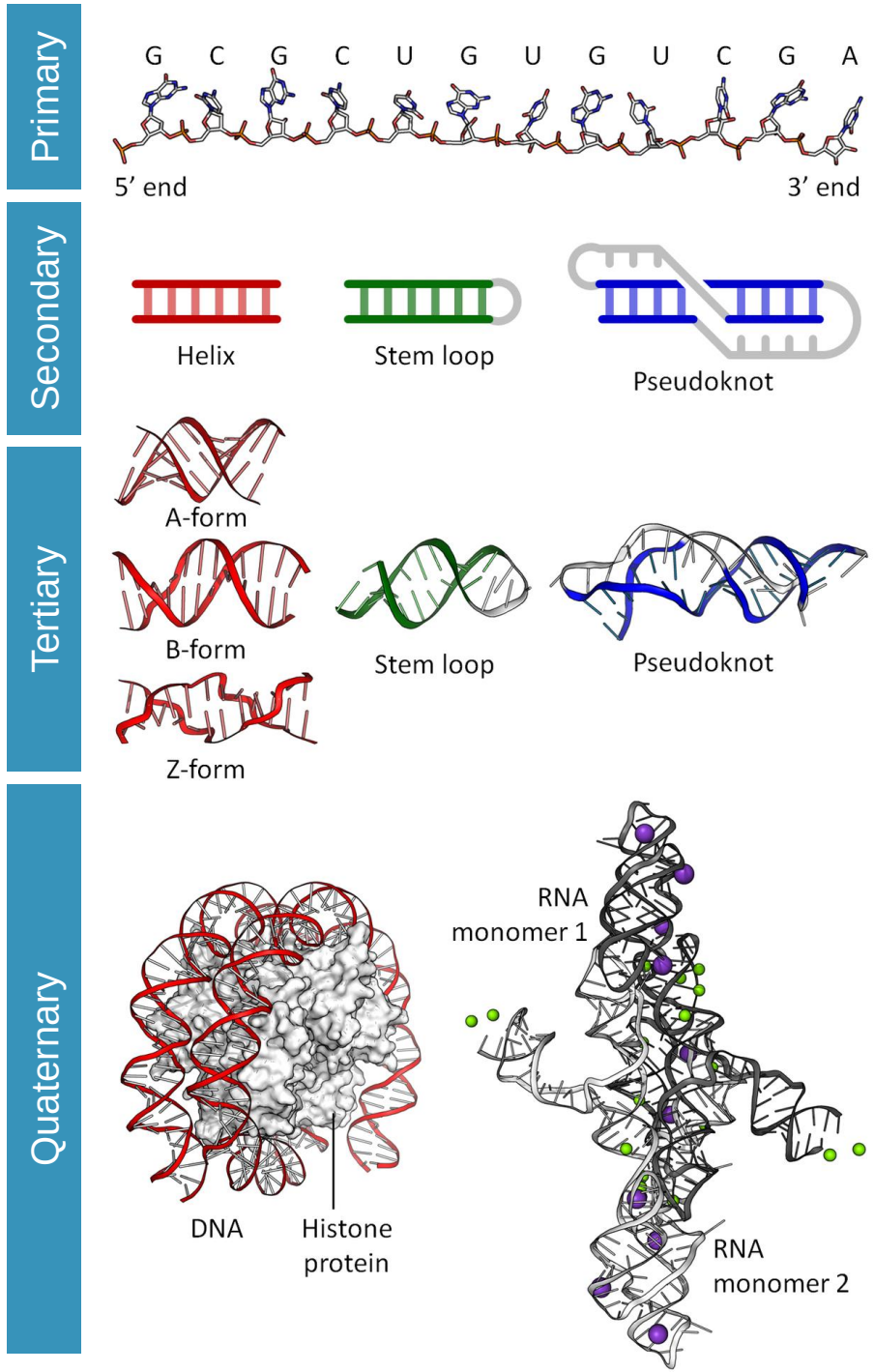
Soluble: Water soluble; can be anywhere in the cell

Protein Domains

- Structural Domain
  - Discrete independently folding unit of a protein
- Conserved Domain (sequence-based)
  - Protein region with recognizable position-specific pattern of sequence conservation
- Sequence-based domains often roughly correspond to structural domains
- Domains often have distinct, identifiable functions

Nucleic Acid Structure

- **Primary-** sequence of nucleotides
- **Secondary-** base pairing interactions between polymers (DNA) or within a single polymer (RNA)
- **Tertiary-** 3D folding pattern
- **Quaternary-** interactions of nucleic acids with other molecules (DNA, RNA, or Protein)



NIH National Library of Medicine
National Center for Biotechnology Information

Experimental techniques

Single crystal X-ray diffraction (SC-XRD)

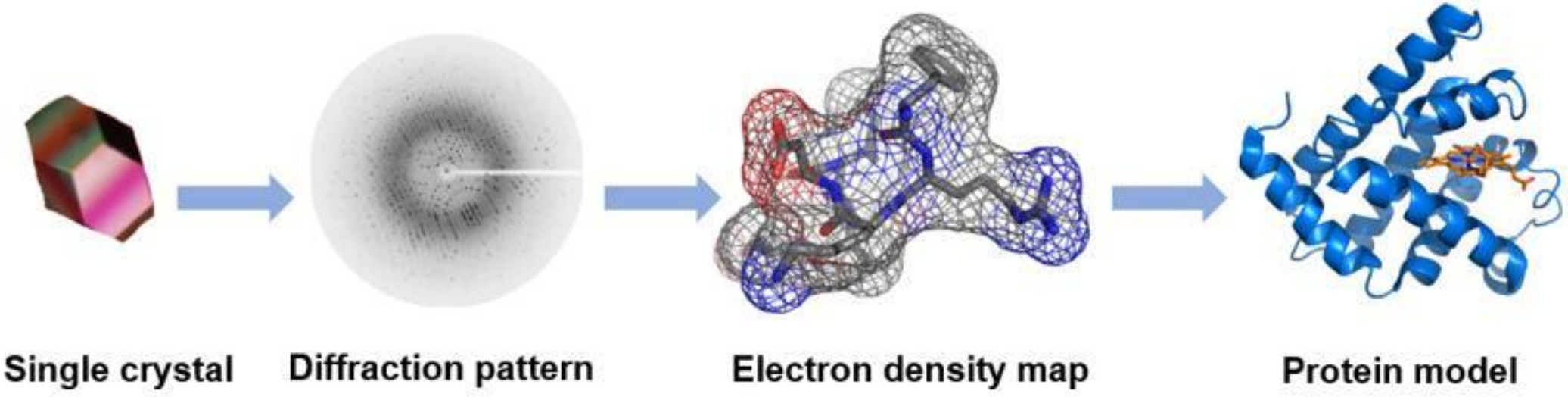
Nuclear magnetic resonance (NMR)

Cryo-electron microscopy (Cryo-EM)

Three main research techniques for structural biology.
According to the statistics of PDB (https://www.rcsb.org/)

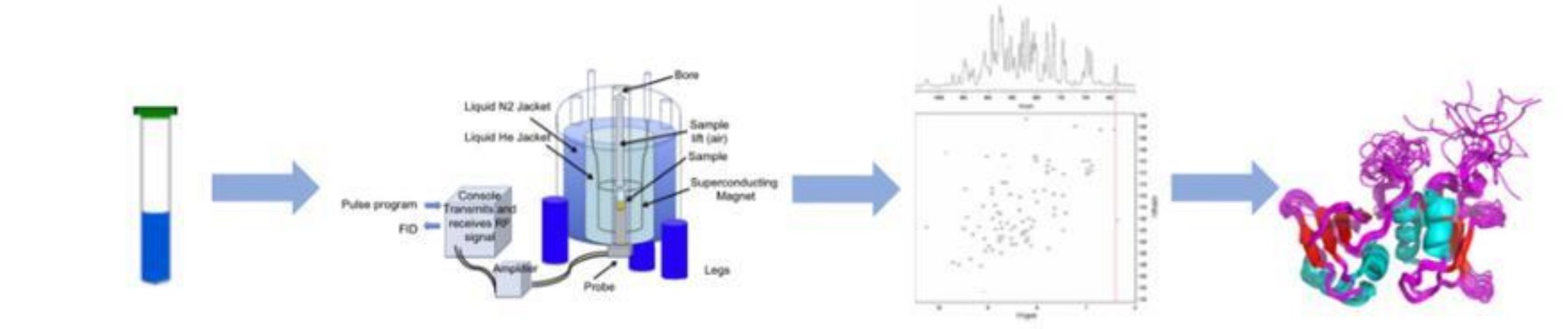
X-ray Crystallography

- Requires crystals, which can be hard to make
- Can handle very large proteins and complexes (e.g. ribosome)
- Provides a “flash picture” with little or no data about motions
- Can include packing artifacts from crystallization



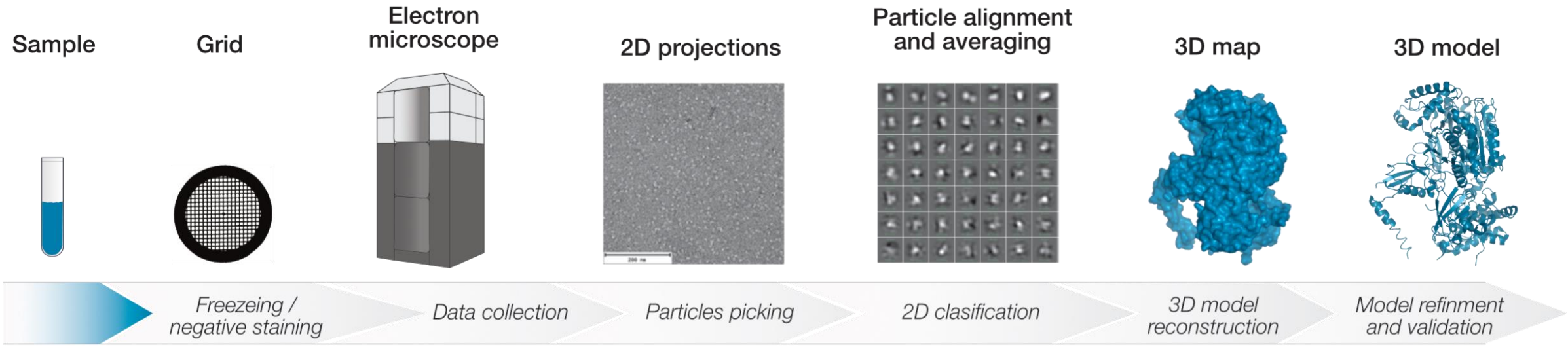
Nuclear Magnetic Resonance

- Requires highly concentrated, C13/N15-labeled protein solutions
- Limited to relatively small proteins (<30 kDa)
- Sensitive to molecular motions
- High protein concentrations may induce non-biological binding



Cryo-electron microscopy

- Requires expensive equipment
- Only small amount of sample
- Rapid freezing sample allows sample to maintain a closer-to-native state
- Useful for biomolecules with high molecular weight

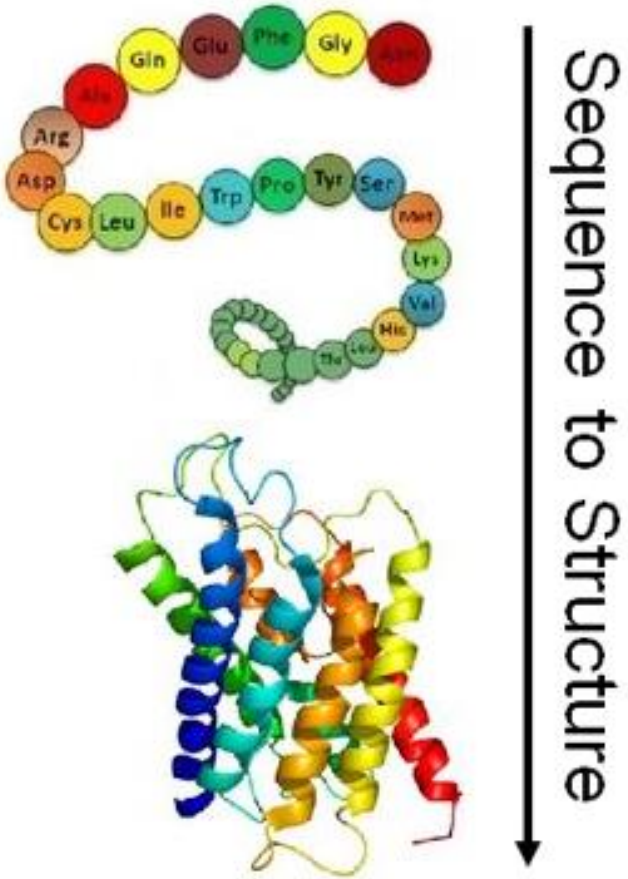


Experimental techniques

	Advantages	Disadvantages
X-ray crystallography	• Well developed • High resolution • Broad molecular weight range	• Difficult sample prep • Static crystalline state
NMR	• High resolution • 3D structure in solution • Good for dynamic study	• Difficult sample prep • High sample purity needed • Static crystalline state captured
Cryo-EM	• Simple sample prep • Structure in native state • Small sample size needed	• Lower resolution • Works best for samples with high molecular weight • EM equipment is costly

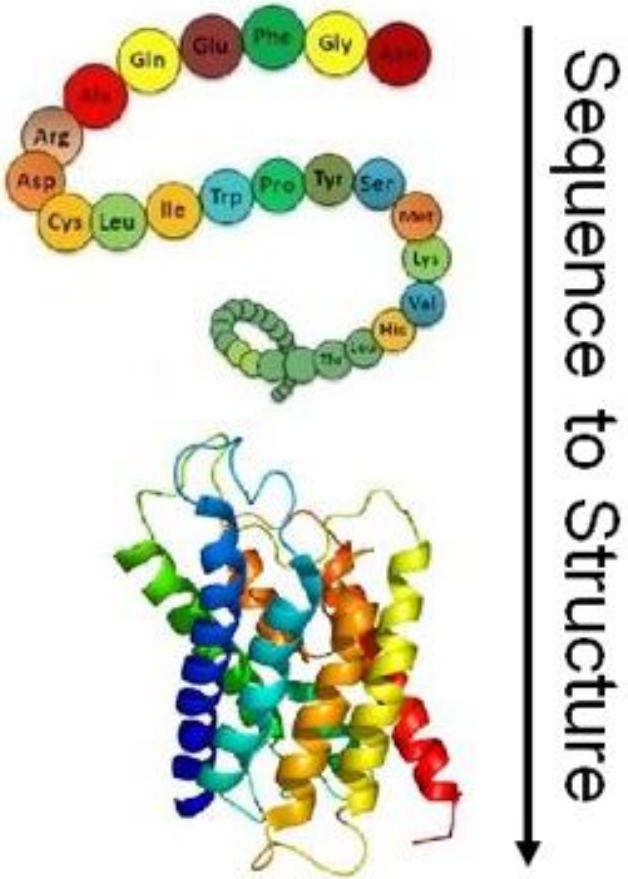
Structure Prediction Overview

- When no or incomplete structure is available
- Prominent research focus of bioinformatics and theoretical chemists
  - Drug discovery
  - Biotechnology/bioengineering
- Predicted structures aren’t housed in most structural databases
  - Structure prediction websites, such as AlphaFold Protein Structure Database, are housed separately



Structure Prediction Methods

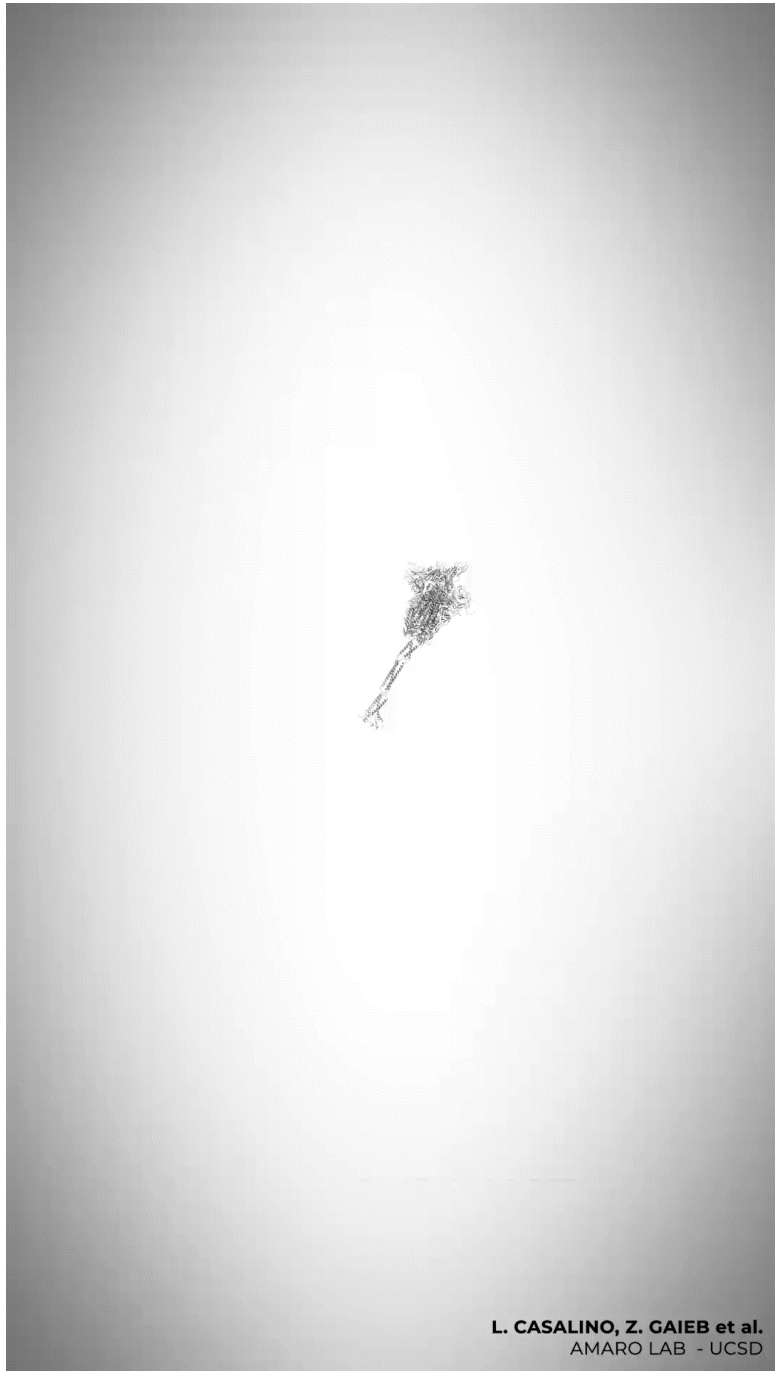
- Comparative Modeling
  - Prediction is based on amino acid sequence and structures of similar molecules available
- Fold recognition
  - Predicts folded structure by aligning a protein of unknown structure and a protein of known structure for low levels of sequence identity (<25%)
- Ab initio
  - Predicts the structure of proteins from the sequence and using molecular energy calculations (Schrodinger equation)



Structure Prediction Example

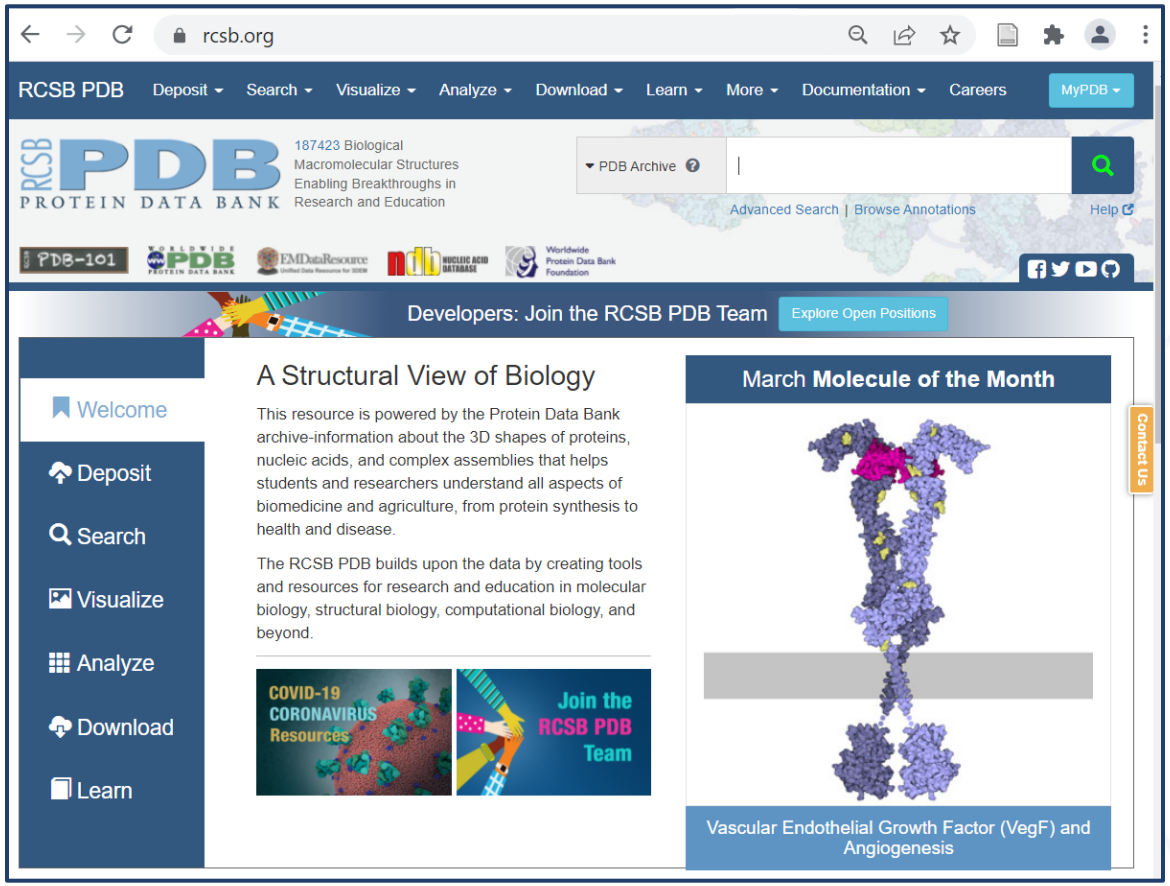
Impact on COVID-19 research

- Researchers have provided key insights into the SARS-CoV-2 proteins through structure prediction
  - Identified critical residues
  - Contextualized variant perturbations
  - Improved understanding of molecular recognition
- Spike fusion glycoprotein example
  - Challenging to characterize experimentally
  - Modeling + molecular dynamics helped researchers understand the roles of glycans on the dynamics of the protein

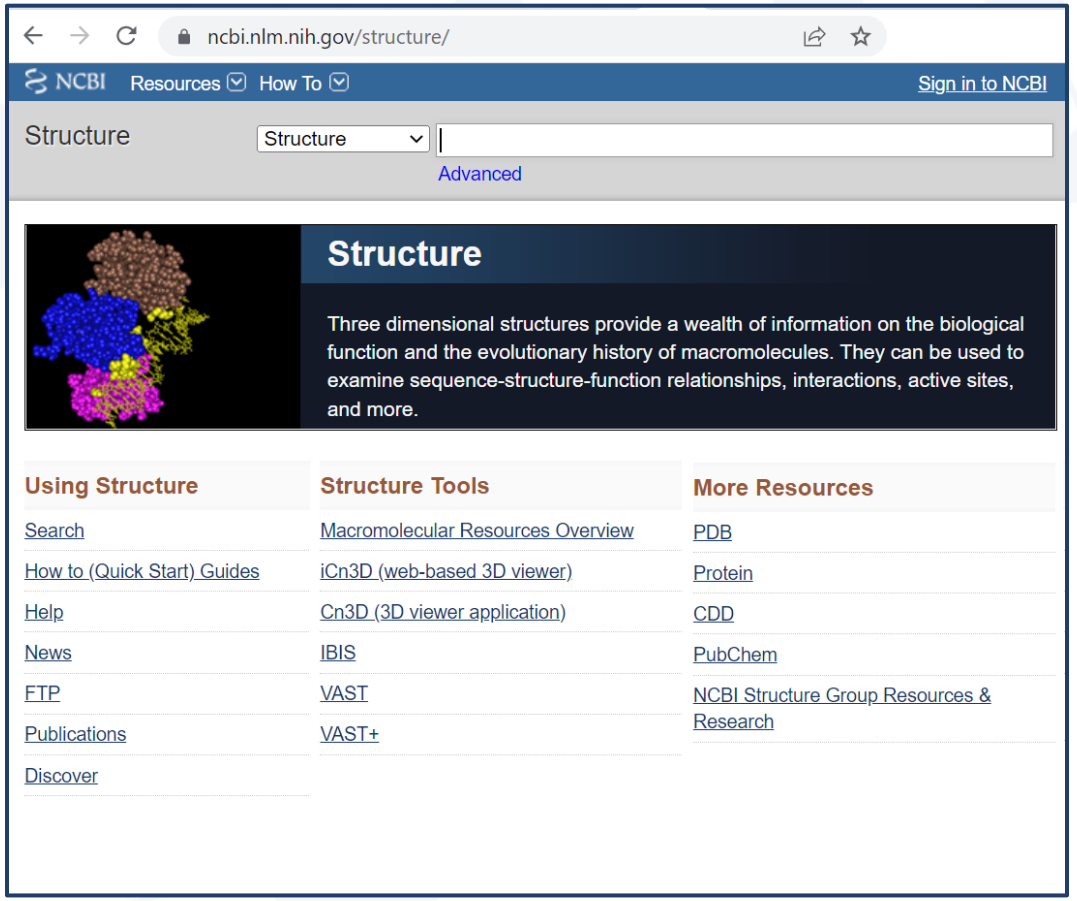


Casalino et al, *Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein*, PMID:33140034

Where do I find structures?



RCSB Protein Data Bank



NCBI Structure Database

Protein Data Bank (PDB)

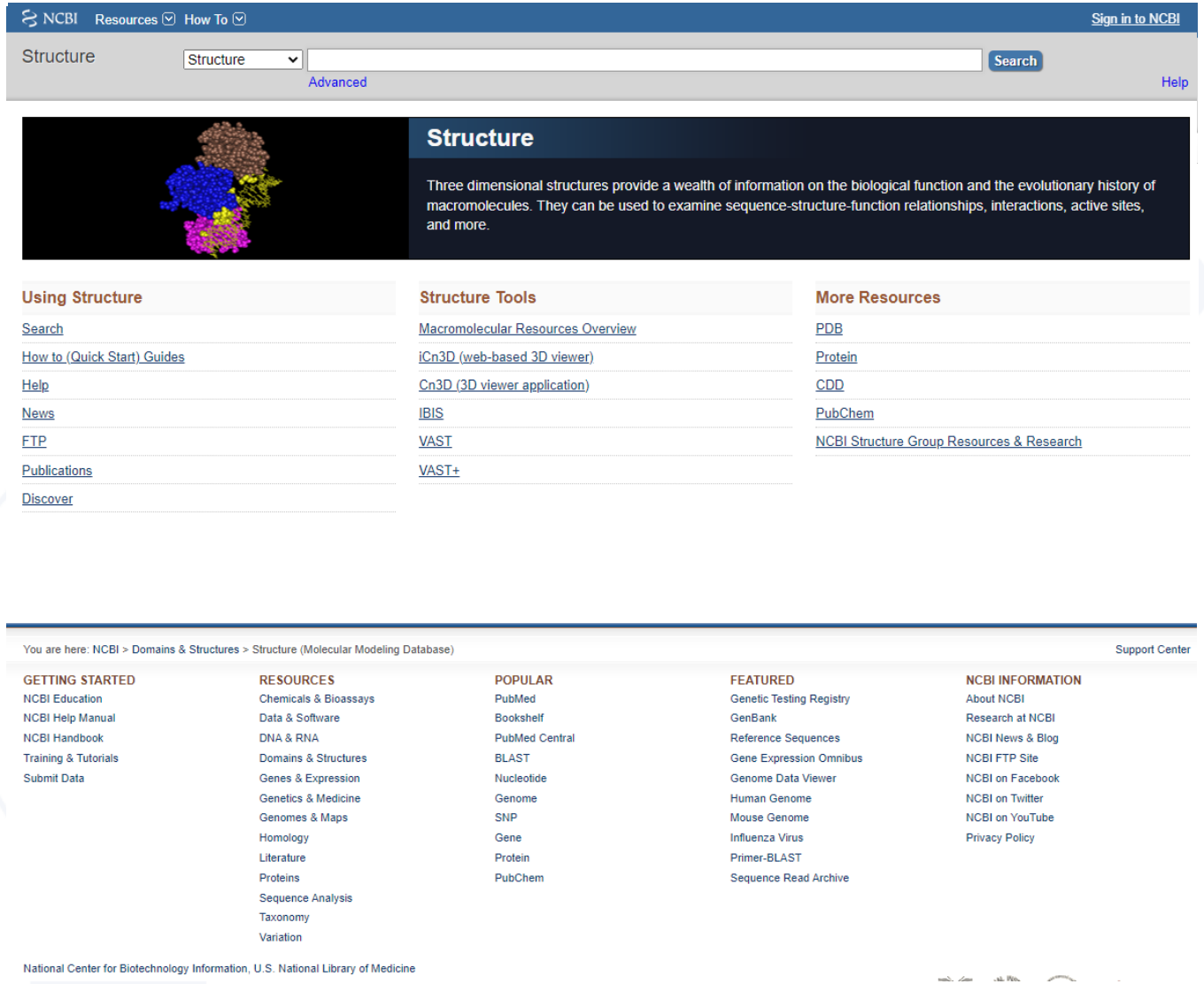
~20 New Structures are deposited daily

Each structure contains:

- 3D atomic coordinates
- Mandatory Metadata
  - Author Information
  - Primary citation
  - Experimental Data
  - Polymer sequence(s)- proteins, DNA, RNA
  - Small Chemical component structures- ligands, inhibitors, etc.

Structure Database

- Updated monthly
- Derived from PDB records
- Additional information added, including:
  - Explicit chemical graph information
  - Validation (secondary structure elements)
  - Includes taxonomy
- Connects 3D to associated literature, molecular data, chemical data, and other NCBI tools



Search Tips

Entrez is a molecular biology database system that provides access to a wealth of NCBI data

- More Entrez Help is available on the NCBI website

Finding structures with Entrez

```
"term1"[field1] AND/OR/NOT "term2"[field2] AND/OR/NOT …
```

- Use field limits and Boolean operators
- Put phrases in quotes

Search Examples

Useful Search Fields
Organism Ex. “Homo sapiens”[orgn]
Experimental Method Ex. “NMR”[exp]
Chemical Name “zinc”[chemical name]
PDB Description Ex. “Tumor Suppressor p53”[title]
[Filter] Ex. “Complex DNA”[filter]

[More Search Field Options]

```
term1[field1] AND/OR/NOT term2[field2] AND/OR/NOT …
```

```
“Homo sapiens”[orgn] AND “X-ray diffraction”[exp]
```

Search results

Items: 1 to 20 of 47803

```
“Homo sapiens”[orgn] AND “X-ray diffraction”[exp]AND “zinc”[chemical name]
```

Search results

Items: 1 to 20 of 6092

```
“Homo sapiens”[orgn] AND “X-ray diffraction”[exp] AND “zinc”[chemical name] AND “Complex DNA”[filter]
```

Search results

Items: 1 to 20 of 288

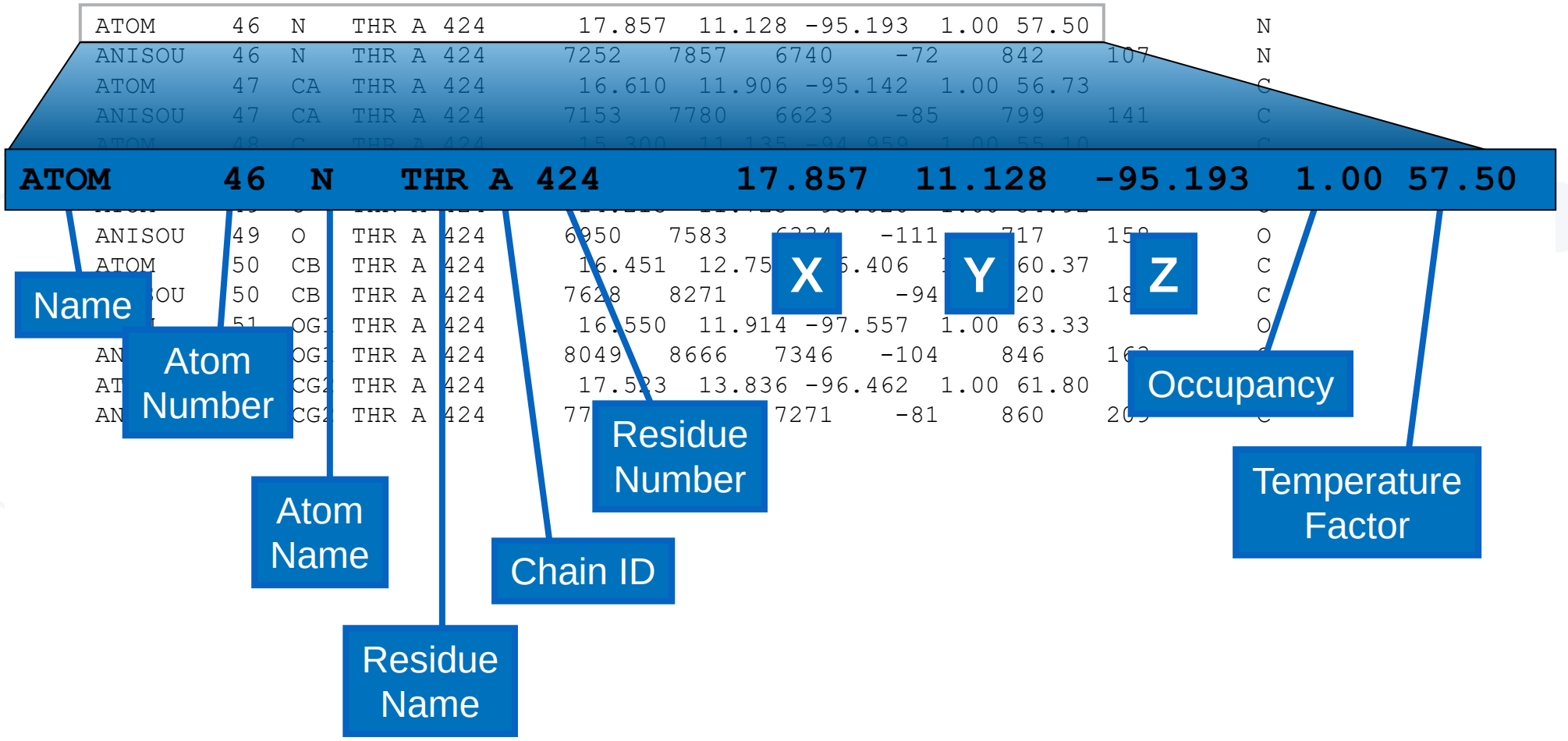
```
1TUP
```

PDB File

```
HEADER    ISOMERASE/DNA                           04-OCT-07   2RGR
TITLE     TOPOISOMERASE IIA BOUND TO G-SEGMENT DNA
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: DNA TOPOISOMERASE 2;
COMPND   3 CHAIN: A;
COMPND   4 FRAGMENT: DNA BINDING AND CLEAVAGE DOMAIN (RESIDUES 419-
COMPND   5 1177);
COMPND   6 SYNONYM: DNA TOPOISOMERASE II;
COMPND   7 EC: 5.99.1.3;
COMPND   8 ENGINEERED: YES;
COMPND   9 MOL_ID: 2;
COMPND  10 MOLECULE: DNA;
COMPND  11 CHAIN: C;
COMPND  12 ENGINEERED: YES;
COMPND  13 MOL_ID: 3;
COMPND  14 MOLECULE: DNA;
COMPND  15 CHAIN: D;
COMPND  16 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: SACCHAROMYCES CEREVISIAE;
SOURCE   3 ORGANISM_COMMON: BAKER'S YEAST;
SOURCE   4 ORGANI
SOURCE   5 GENE:
SOURCE   6 EXPRES
SOURCE   7 EXPRES
SOURCE   8 EXPRES
SOURCE   9 EXPRES
SOURCE  10 EXPRES
SOURCE  11 EXPRES
SOURCE  12 MOL_ID
SOURCE  13 SYNTHE
SOURCE  14 MOL_ID
SOURCE  15 SYNTHE
```

```
REMARK   2
REMARK   2 RESOLUTION.    3.00 ANGSTROMS.
REMARK   3
REMARK   3 REFINEMENT.
REMARK   3   PROGRAM     : PHENIX
…
REMARK 280
REMARK 280 CRYSTAL
REMARK 280 SOLVENT CONTENT, VS   (%): 59.90
REMARK 280 MATTHEWS COEFFICIENT, VM (ANGSTROMS**3/DA): 3.07
REMARK 280
REMARK 280 CRYSTALLIZATION CONDITIONS: 12-20% PEG 1000, 100-250 MM MGCL2,
REMARK 280  100 MM SODIUM CACODYLATE, PH 7.0, VAPOR DIFFUSION, HANGING
REMARK 280  DROP, TEMPERATURE 277K
REMARK 290
```

PDB File: Data



More about PDB Data

Computational Structural Biology

- Structure Prediction- inference of 3D structure from sequence data
- Molecular Docking- predicts the orientation of one molecule to another
- Molecular Dynamics Simulations- analyzes physical movements of atoms and molecules over time

NCBI

Computational Structural Biology

- Structure Prediction- inference of 3D structure from sequence data
- Molecular Docking- predicts the orientation of one molecule to another
- Molecular Dynamics Simulations- analyzes physical movements of atoms and molecules over time

- Rely on experimental information from public databases
  - NCBI Databases and RCSB Protein Data Bank

NCBI

Homology Modeling *vs Ab initio* Prediction

Ab initio Prediction	**Comparative Modeling**
Applicable to any sequence	Applicable to only those sequences with recognizable similarity to a template structure
Not very accurate (>4Å RMSD)	Fairly accurate (<3Å RMSD), similar to low resolution X-ray structure
Attempted for proteins of <100 residues	Not limited by size
Accuracy and applicability are limited by our understanding of the protein folding problem	Accuracy and applicability are limited by the number of known folds