BIMM 143

Structural Bioinformatics

Lecture 11

Barry Grant

UC San Diego

http://thegrantlab.org/bimm143

http://www.ks.uiuc.edu/Development/Download/download.cgi

"*Bioinformatics is the application of computers to the collection, archiving, organization, and analysis of biological data.*"

… A hybrid of biology and computer science

"*Bioinformatics is the application of computers to the collection, archiving, organization, and analysis of biological data.*"

Bioinformatics is computer aided biology!

"*Bioinformatics is the application of computers to the collection, archiving, organization, and analysis of biological data.*"

Bioinformatics is computer aided biology!

Goal: Data to Knowledge

So what is **structural bioinformatics**?

So what is **structural bioinformatics**?

… computer aided structural biology!

Aims to characterize and interpret biomolecules and their assembles at the molecular & atomic level

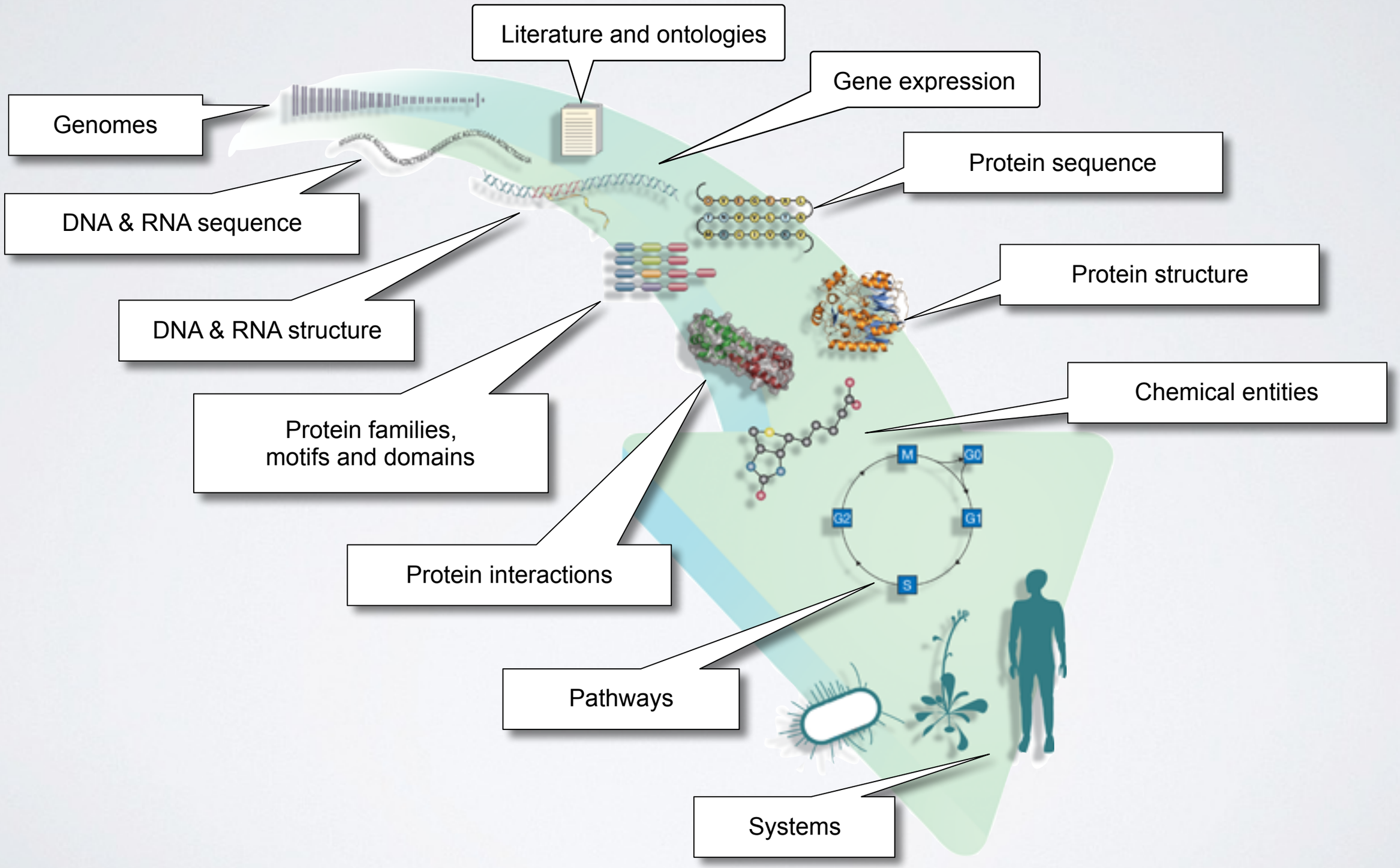
Why should we care?

Why should we care?

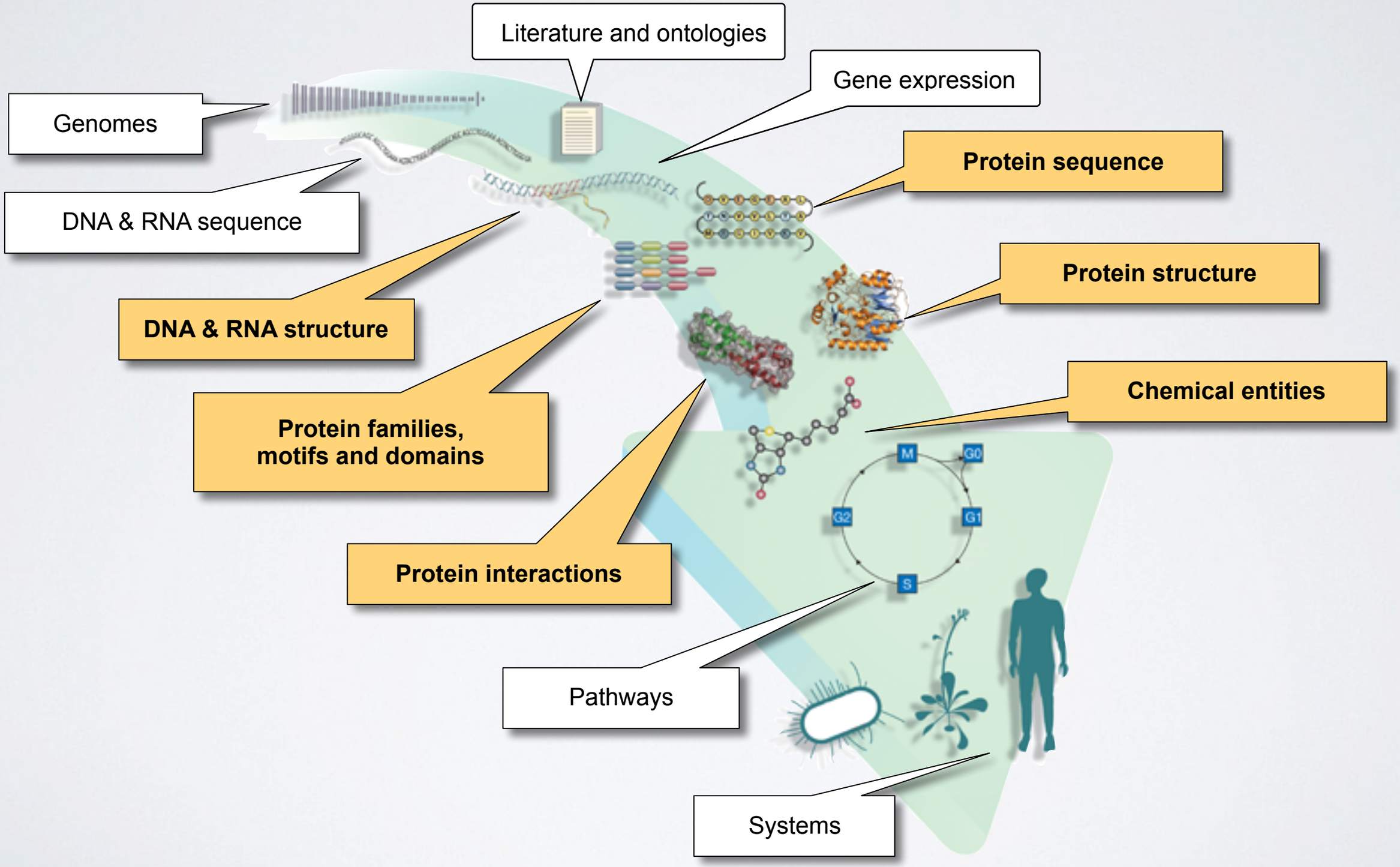
Because biomolecules are “nature’s robots”

… and because it is only by coiling into **specific 3D structures** that they are able to perform their functions

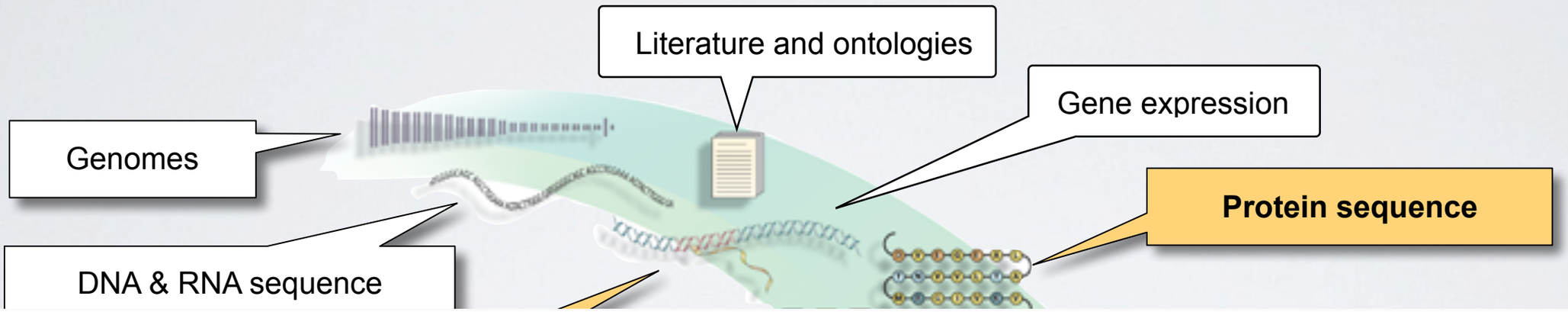
BIOINFORMATICS DATA



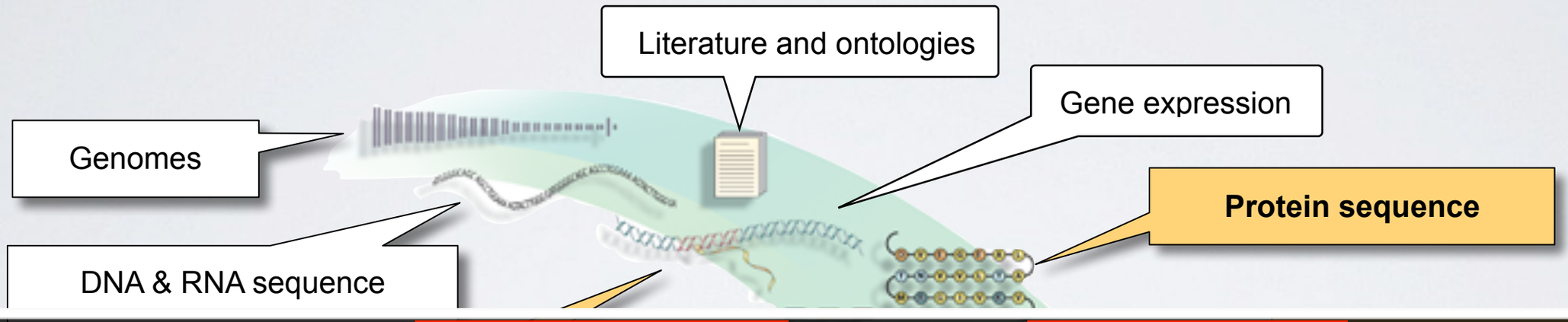
STRUCTURAL DATA IS CENTRAL

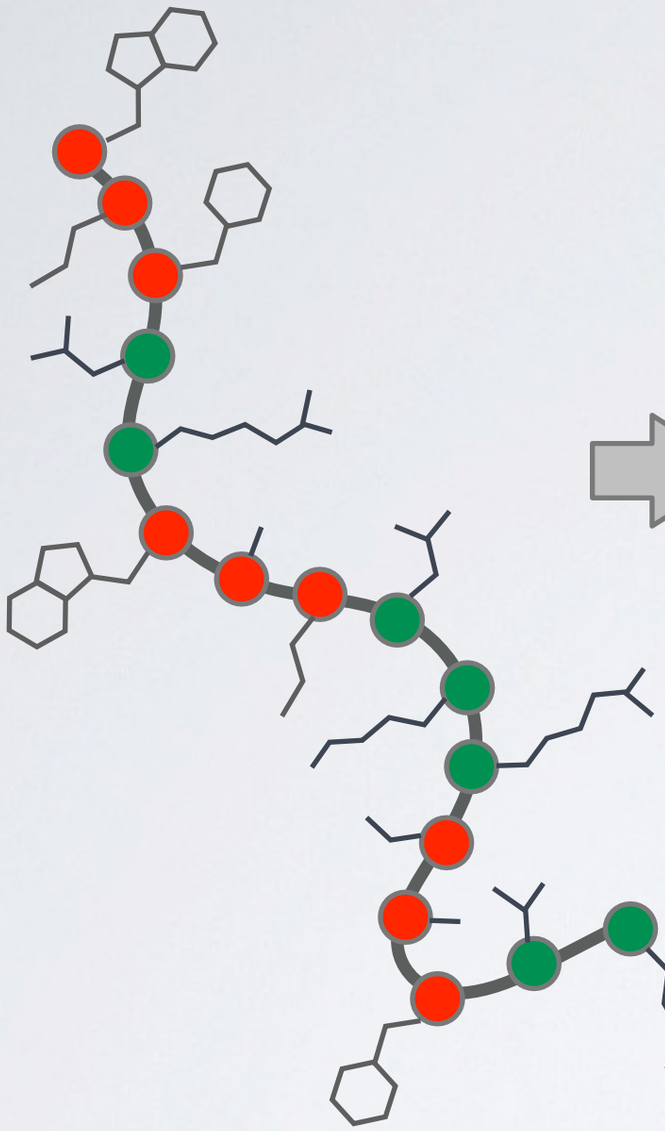


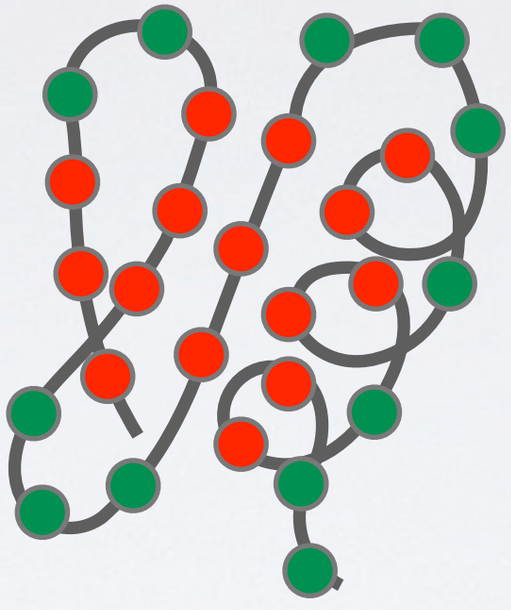
STRUCTURAL DATA IS CENTRAL

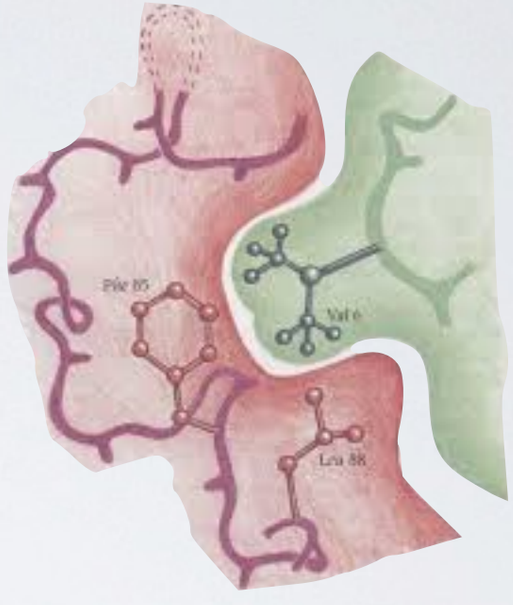


STRUCTURAL DATA IS CENTRAL









Sequence	Structure	Function
• Unfolded chain of amino acid chain • Highly mobile • Inactive	• Ordered in a precise 3D arrangment • Stable but dynamic	• Active in specific “conformations” • Specific associations & precise reactions

In daily life, we use machines with functional *structure* and *moving parts*

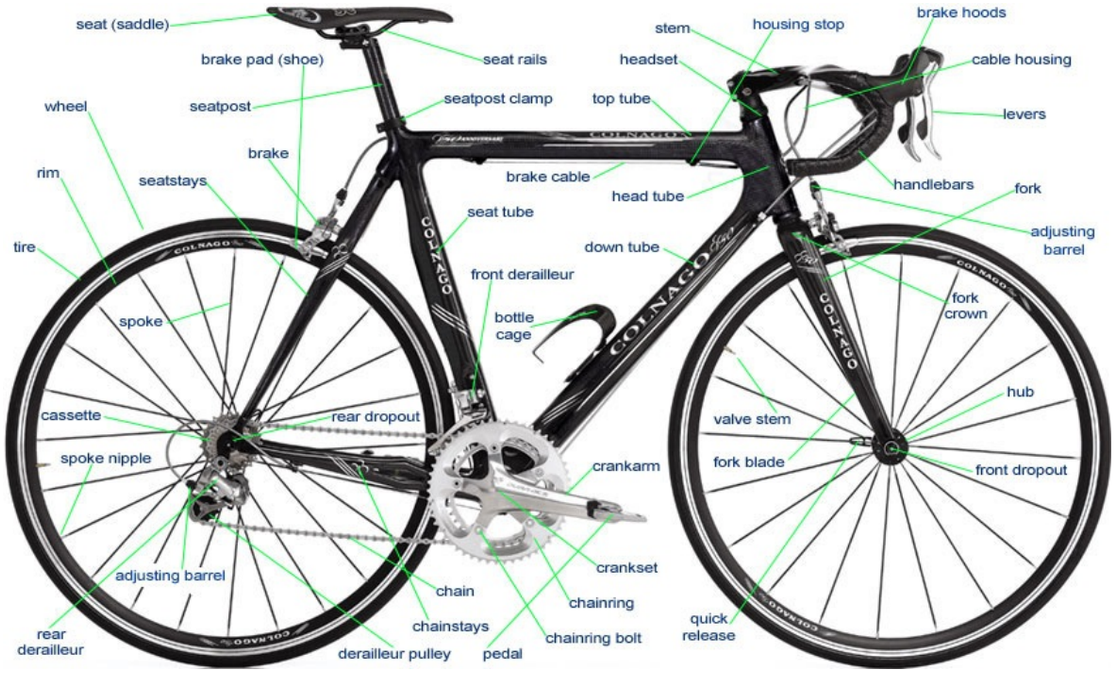
Genomics is a great start ….

Track Bike – DL 175

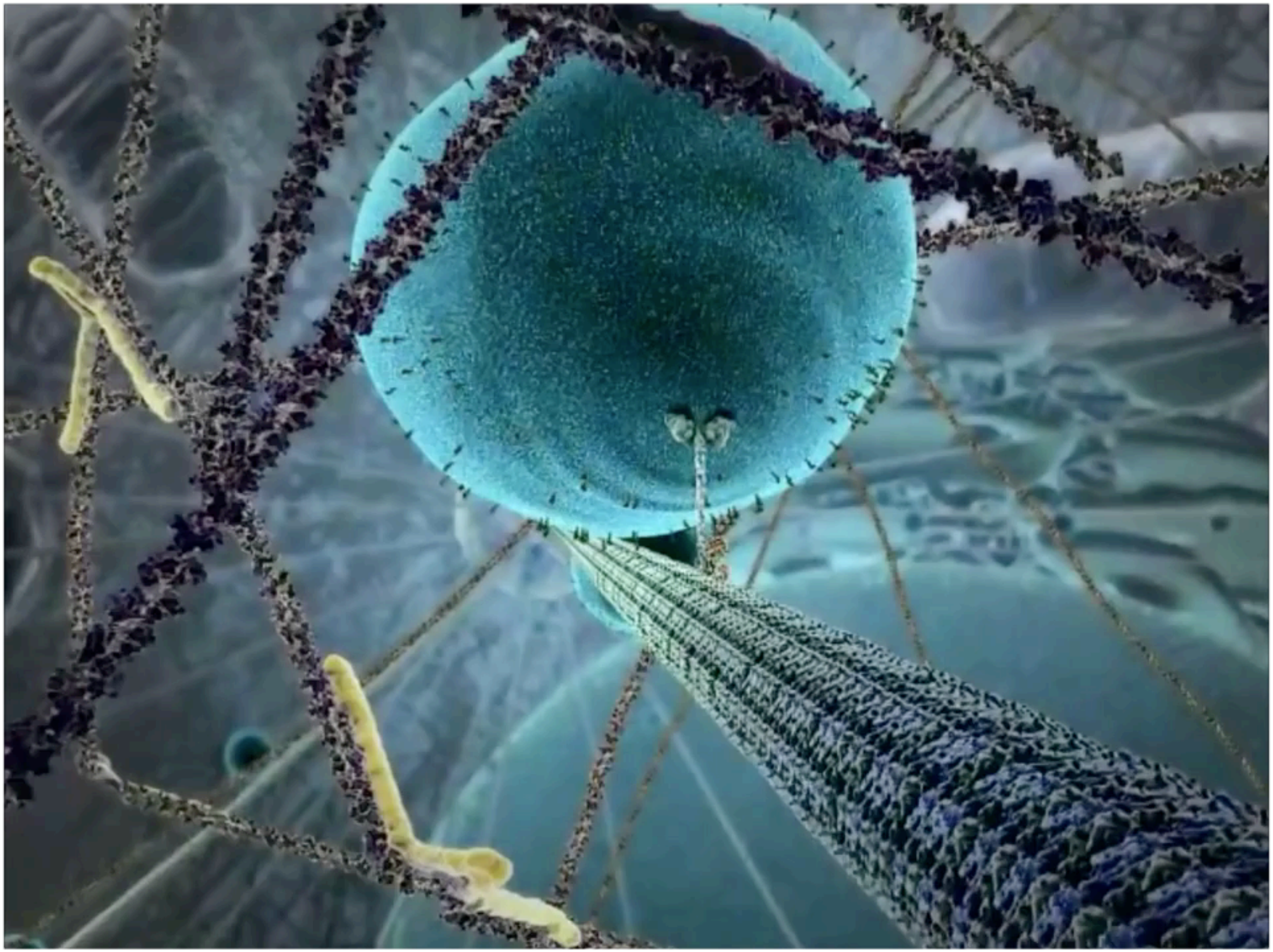
REF. NO.	IBM NO.	DESCRIPTION
1	156011	Track Frame 21", 22", 23", 24", Team Red
2	157040	Fork for 21" Frame
2	157039	Fork for 22" Frame
2	157038	Fork for 23" Frame
2	157037	Fork for 24" Frame
3	191202	Handlebar TTT Competition Track Alloy 15/16"
4		Handlebar Stem, TTT, Specify extension
5	191278	Expander Bolt
6	191272	Clamp Bolt
7	145841	Headset Complete 1 x 24 BSC
8	145842	Ball Bearings
9	190420	175 Raleigh Pistard Seta Tubular Prestavalve 27"
10	190233	Rim, 27" AVA Competition (36H) Alloy Prestavalve
11	145973	Hub, Large Flange Campagnolo Pista Track Alloy (pairs)
12	190014	Spokes, 11 5/8"
13	145837	Sleeve
14	145636	Ball Bearings
15	145170	Bottom Bracket Axle
16	145838	Cone for Sleeve
17	146473	L.H. Adjustable Cup
18	145833	Lockring
19	145239	Straps for Toe Clips
20	145834	Fixing Bolt
21	145835	Fixing Washer
22	145822	Dustcap
23	145823	R.H. and L.H. Crankset with Chainwheel
24	146472	Fixed Cup
25	145235	Toe Clips, Christophe, Chrome (Medium)
26	145684	Pedals, Extra Light, Pairs
27	123021	Chain
28	145980	Seat Post
29		Seat Post Bolt and Nut
30	167002	Saddle, Brooks
31	145933	Track Sprocket, Specify 12, 13, 14, 15, or 16 T.

- But a parts list is not enough to understand how a bicycle works

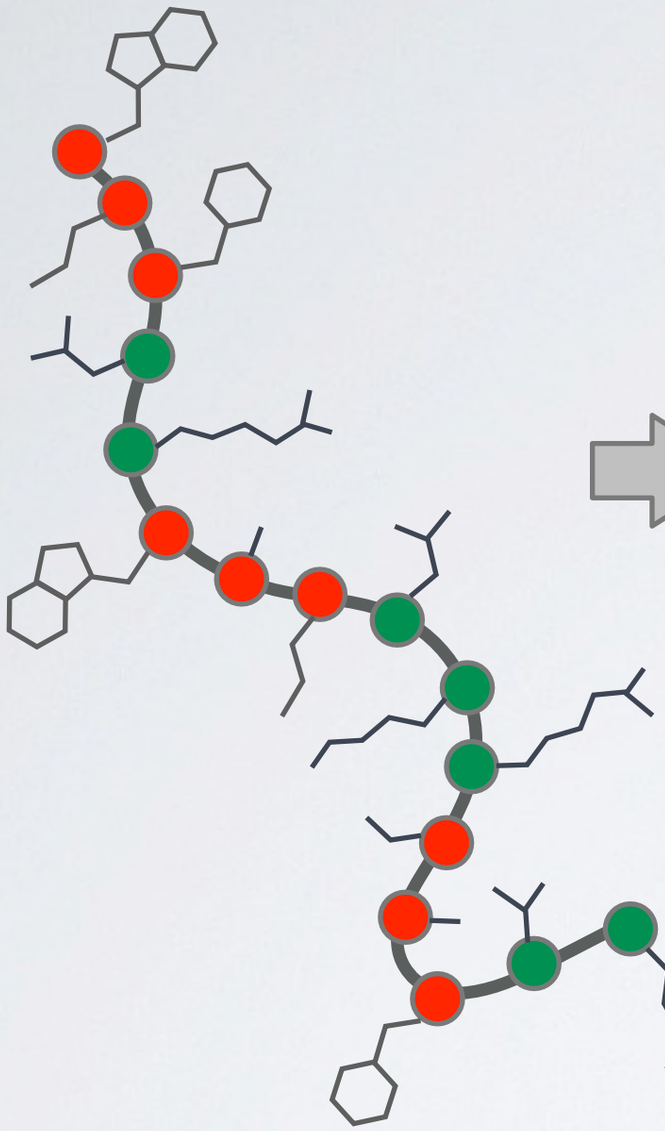
… but not the end

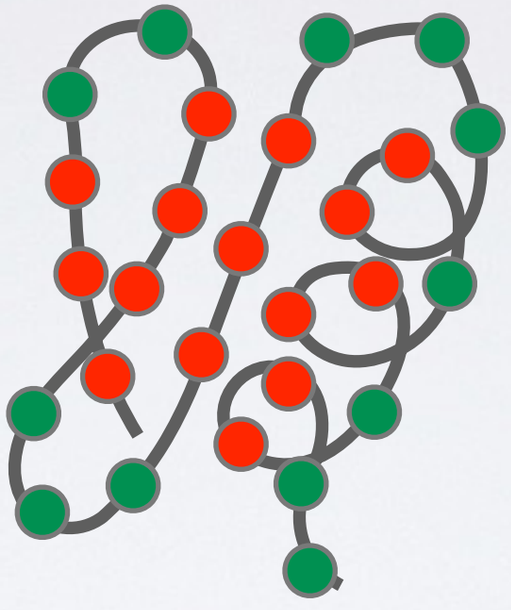


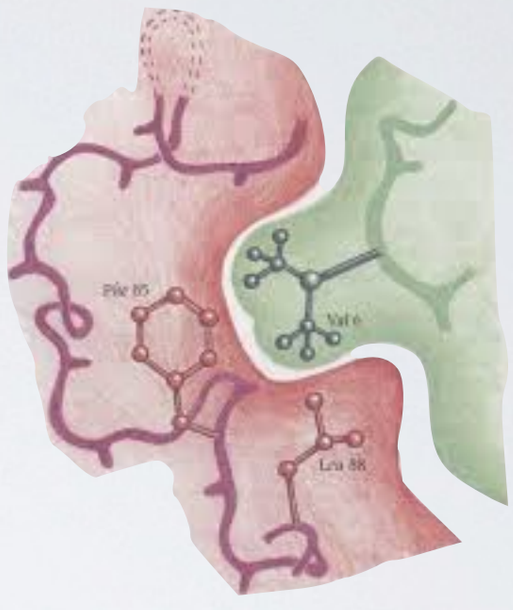
- We want the full spatiotemporal picture, and an ability to control it
- Broad applications, including drug design, medical diagnostics, chemical manufacturing, and energy



Extracted from The Inner Life of a Cell by Cellular Visions and Harvard
[YouTube link: https://www.youtube.com/watch?v=y-uuk4Pr2i8]

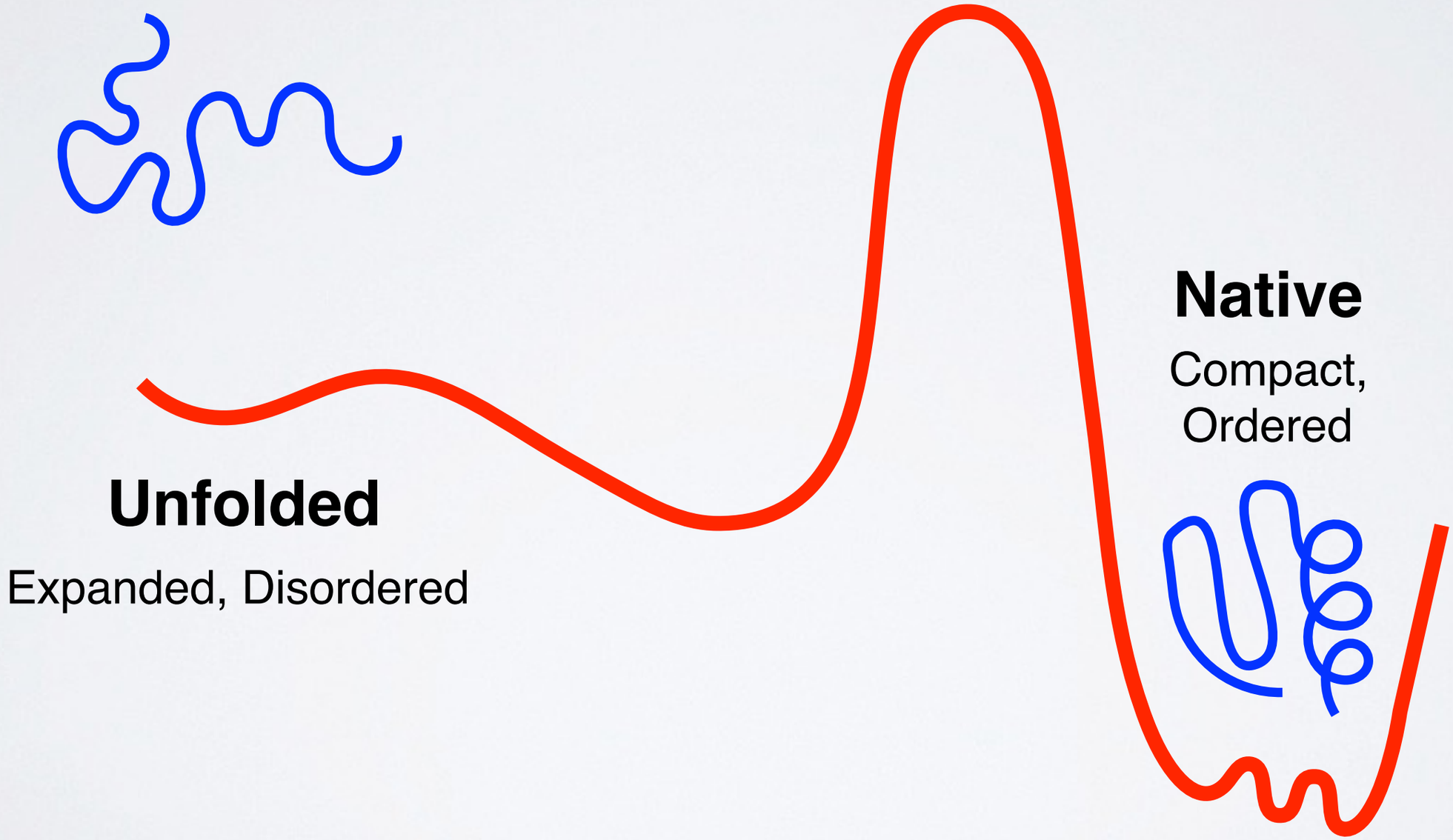






Sequence	Structure	Function
• Unfolded chain of amino acid chain • Highly mobile • Inactive	• Ordered in a precise 3D arrangment • Stable but dynamic	• Active in specific "conformations" • Specific associations & precise reactions

KEY CONCEPT: ENERGY LANDSCAPE



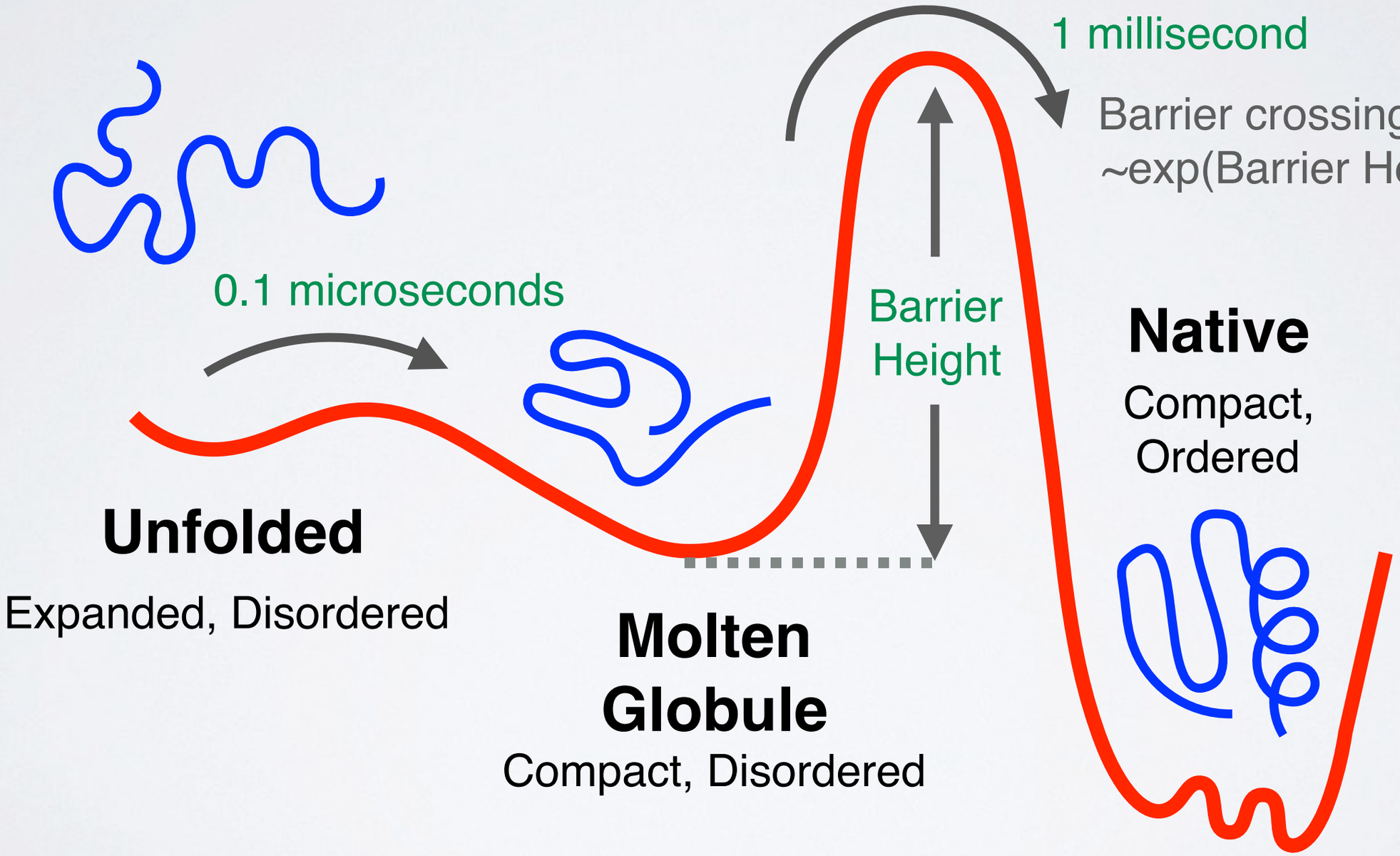
Unfolded

Expanded, Disordered

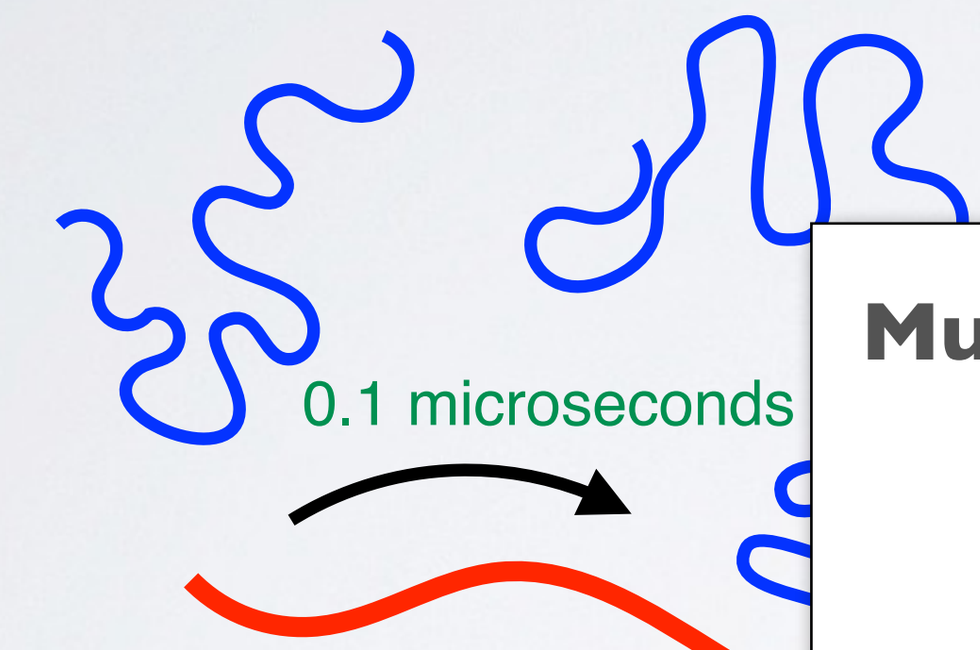
Native

Compact,
Ordered

KEY CONCEPT: ENERGY LANDSCAPE



KEY CONCEPT: **ENERGY LANDSCAPE**



0.1 microseconds

1 millisecond

Barrier crossing time

Unfolded State

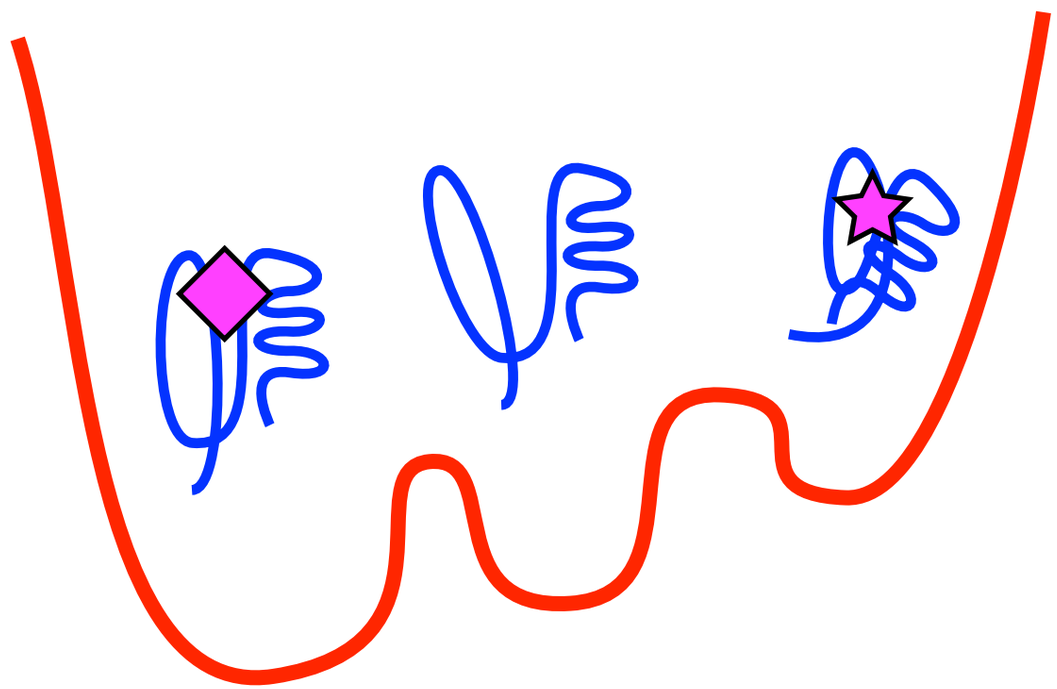
Expanded, Disordered

Mo

Cor

Multiple Native Conformations

(e.g. ligand bound and unbound)



Today's Menu

- **Overview of structural bioinformatics**
  - Motivations, goals and challenges

- **Fundamentals of protein structure**
  - Structure composition, form and forces

- **Representing, interpreting & modeling protein structure**
  - Visualizing & interpreting protein structures
  - Analyzing protein structures
  - Modeling energy as a function of structure

Today's Menu

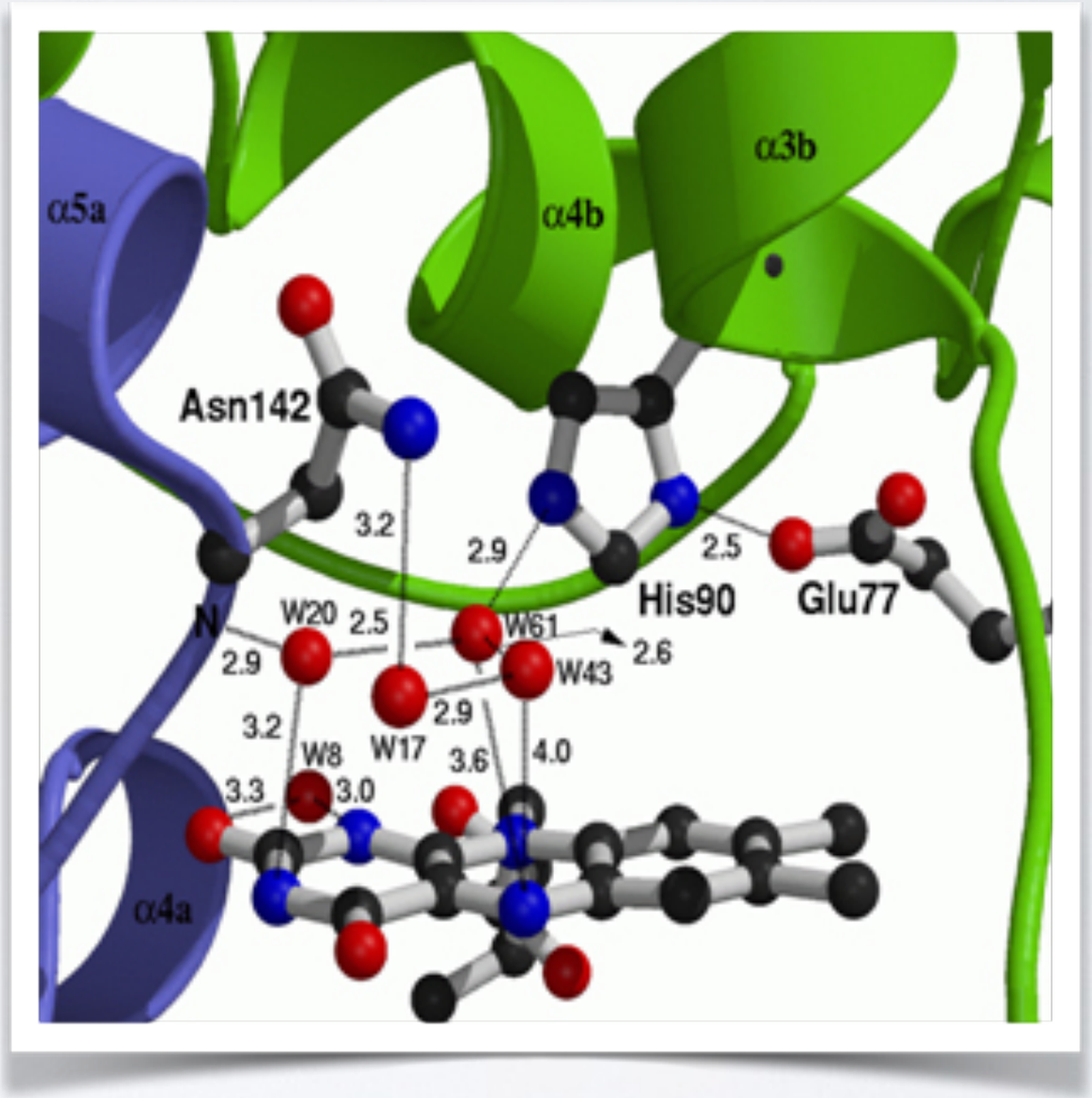
- **Overview of structural bioinformatics**
  - Motivations, goals and challenges

- **Fundamentals of protein structure**
  - Structure composition, form and forces

- **Representing, interpreting & modeling protein structure**
  - Visualizing & interpreting protein structures
  - Analyzing protein structures
  - Modeling energy as a function of structure

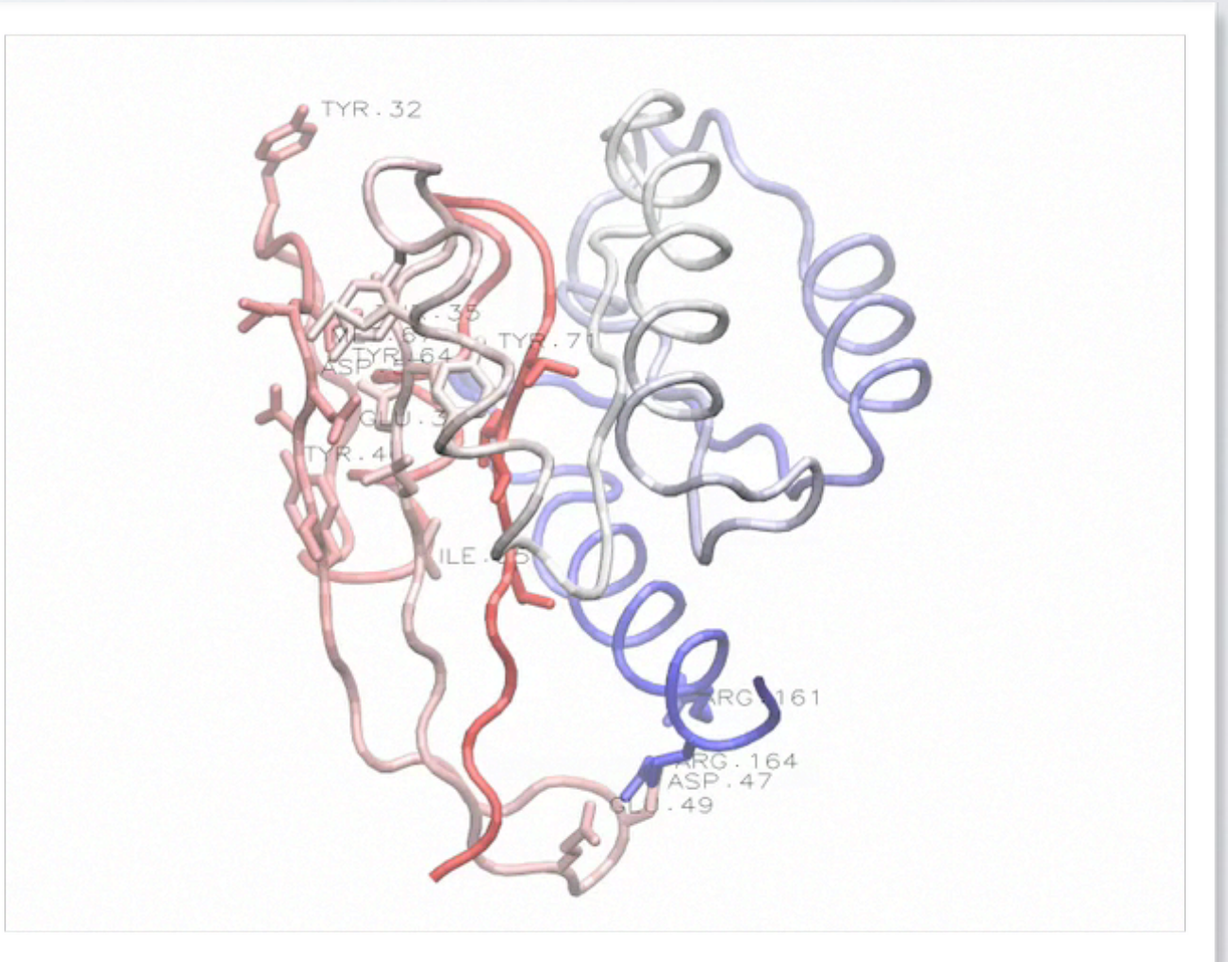
Motivation 1:
Detailed understanding of molecular interactions

Provides an invaluable structural context for conservation and mechanistic analysis leading to functional insight.



Motivation 1:
Detailed understanding of molecular interactions

Computational modeling can provide detailed insight into functional interactions, their regulation and potential consequences of perturbation.

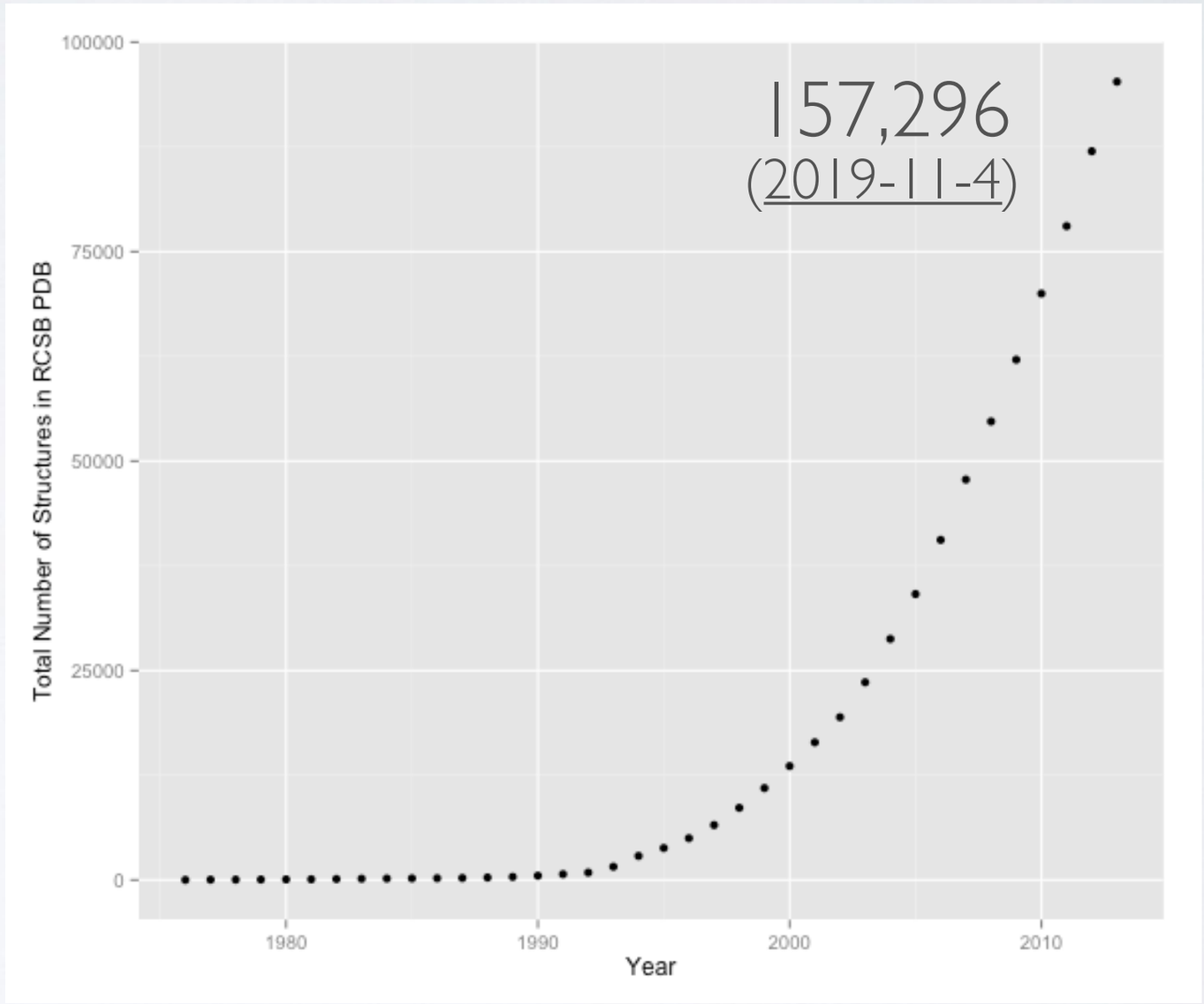


Grant *et al.* PLoS. Comp. Biol. (2010)

Motivation 2:

Lots of structural data is becoming available

Structural Genomics has contributed to driving down the cost and time required for structural determination



Data from: https://www.rcsb.org/stats/

Motivation 2:

Lots of structural data is becoming available

Structural Genomics has contributed to driving down the cost and time required for structural determination

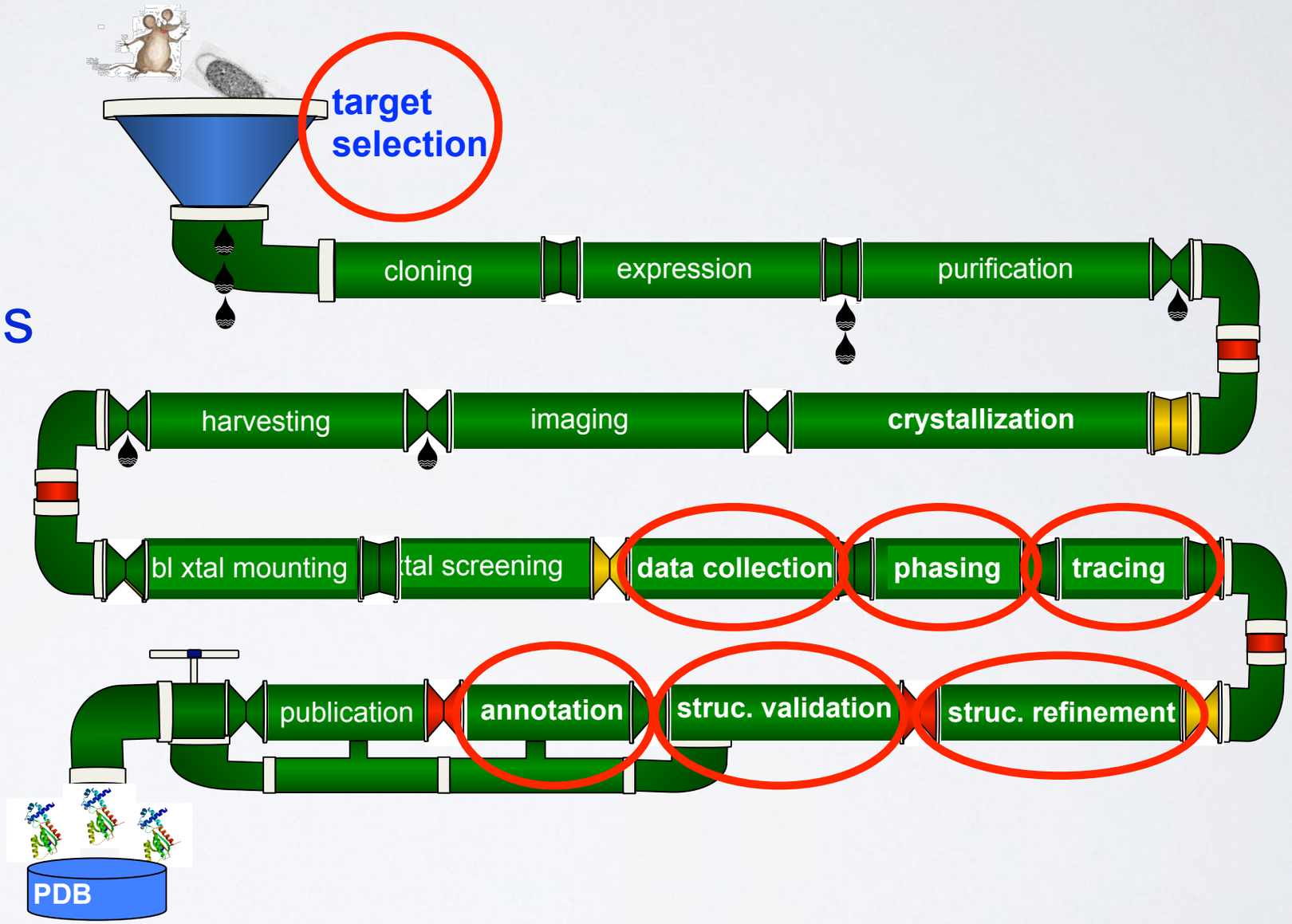
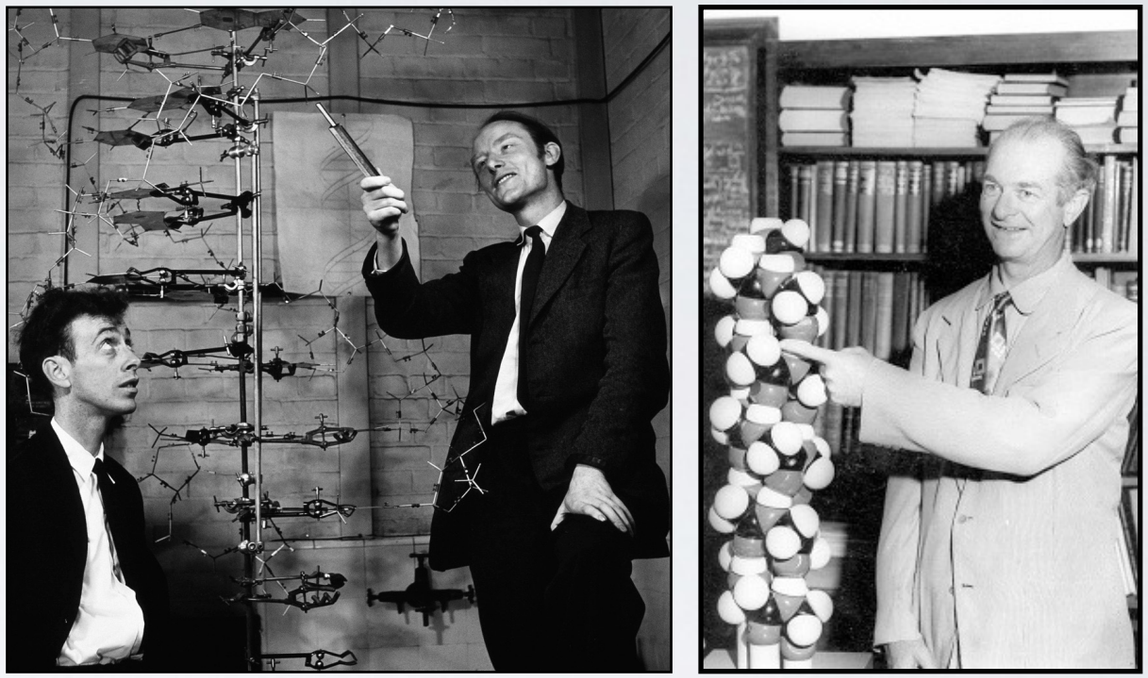


Image Credit: "Structure determination assembly line" Adam Godzik

Motivation 3:

Theoretical and computational predictions have been, and continue to be, enormously valuable and influential!



SUMMARY OF KEY **MOTIVATIONS**

Sequence > Structure > Function

- Structure determines function, so understanding structure helps our understanding of function

Structure is more conserved than sequence

- Structure allows identification of more distant evolutionary relationships

Structure is encoded in sequence

- Understanding the determinants of structure allows design and manipulation of proteins for industrial and medical advantage

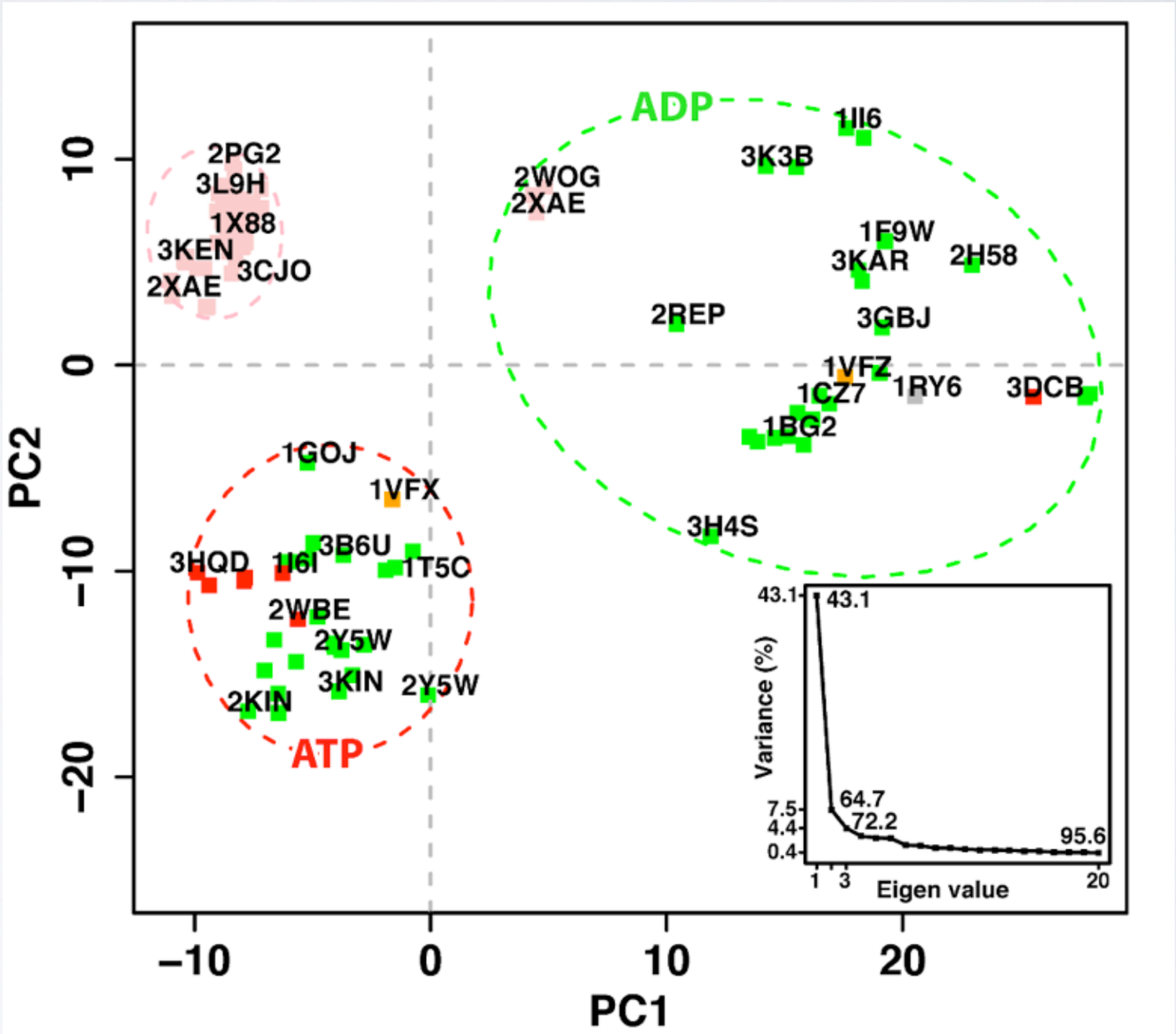
Goals:

- Visualization
- Analysis
- Comparison
- Prediction
- Design

Scarabelli and Grant. PLoS. Comp. Biol. (2013)

Goals:

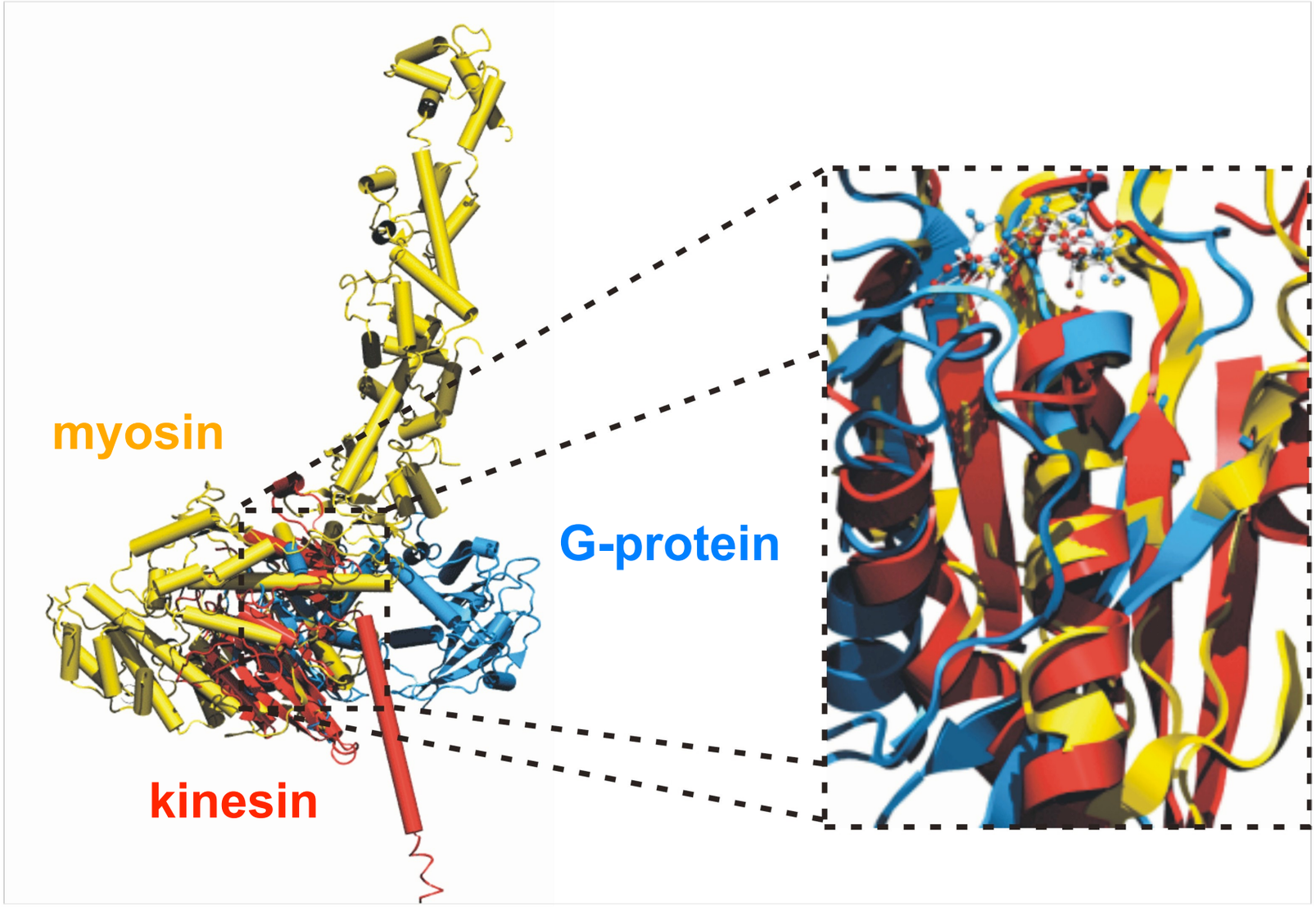
- Visualization
- Analysis
- Comparison
- Prediction
- Design



Scarabelli and Grant. PLoS. Comp. Biol. (2013)

Goals:

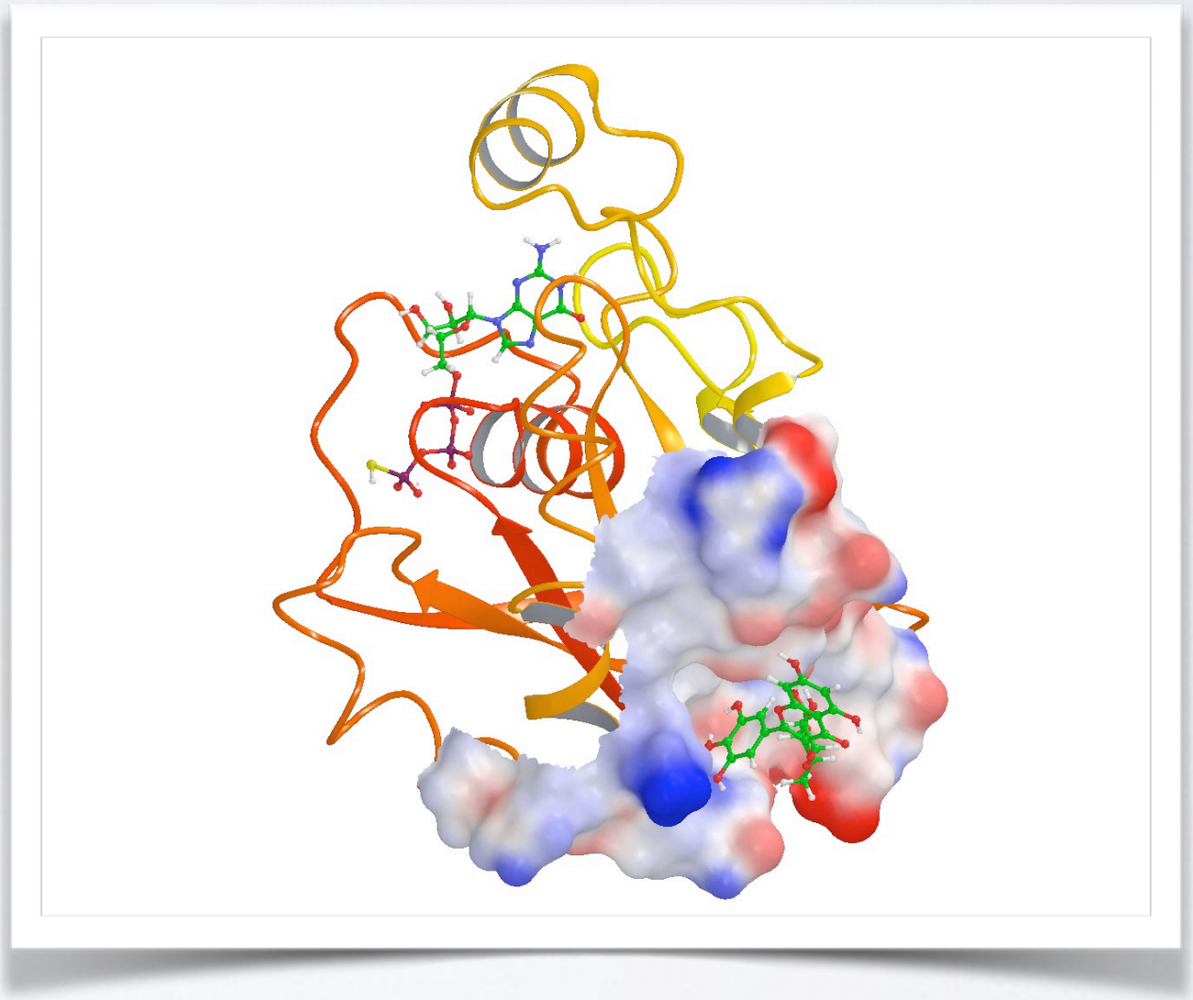
- Visualization
- Analysis
- Comparison
- Prediction
- Design



Grant *et al.* unpublished

Goals:

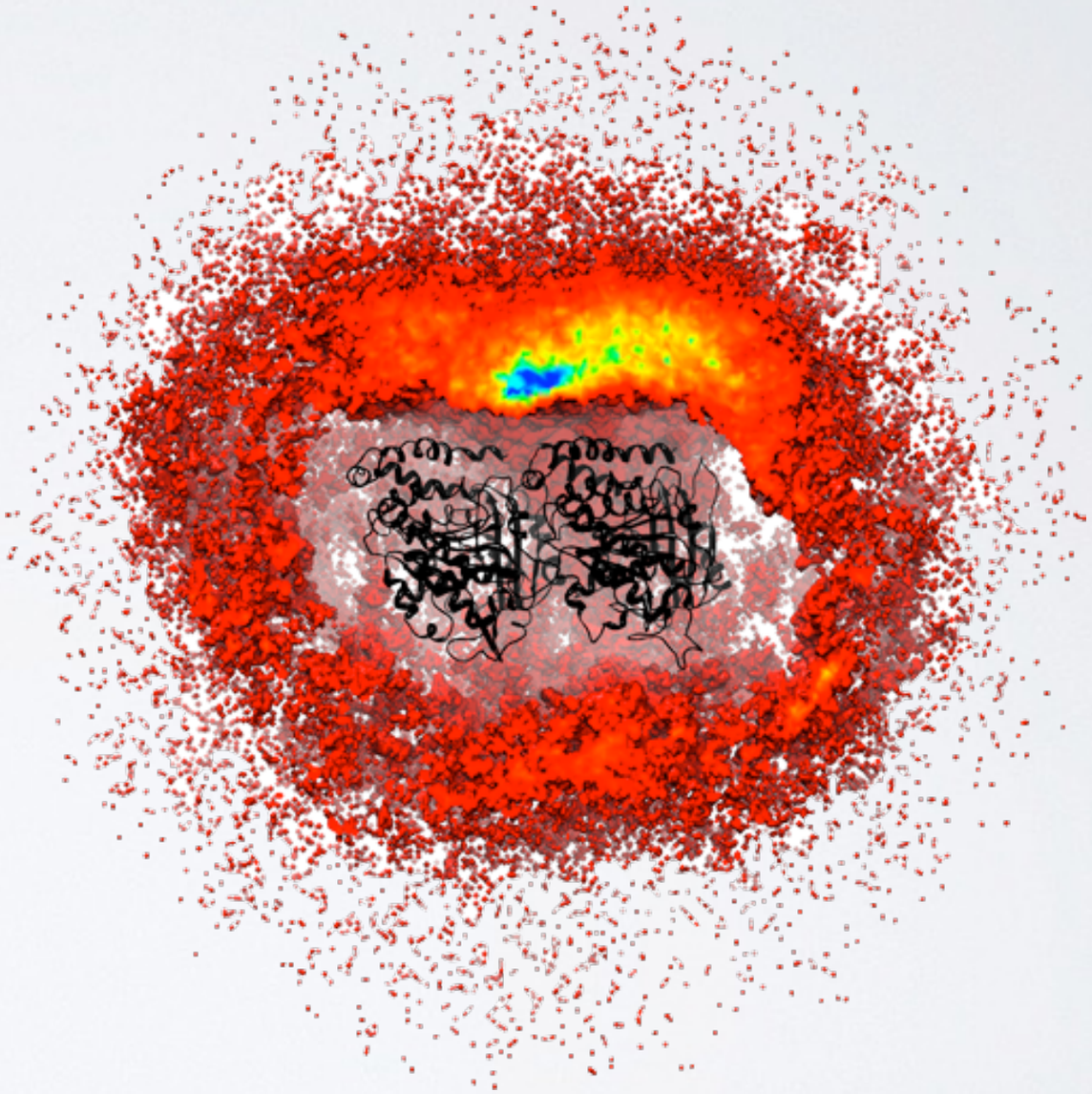
- Visualization
- Analysis
- Comparison
- Prediction
- Design



Grant *et al.* PLoS One (2011, 2012)

Goals:

- Visualization
- Analysis
- Comparison
- Prediction
- Design



Grant *et al.* PLoS Biology (2011)

MAJOR RESEARCH AREAS AND CHALLENGES

Include but are not limited to:

- Protein classification
- Structure prediction from sequence
- Binding site detection
- Binding prediction and drug design
- Modeling molecular motions
- Predicting physical properties (stability, binding affinities)
- Design of structure and function
- etc...

With applications to Biology, Medicine, Agriculture and Industry

Today's Menu

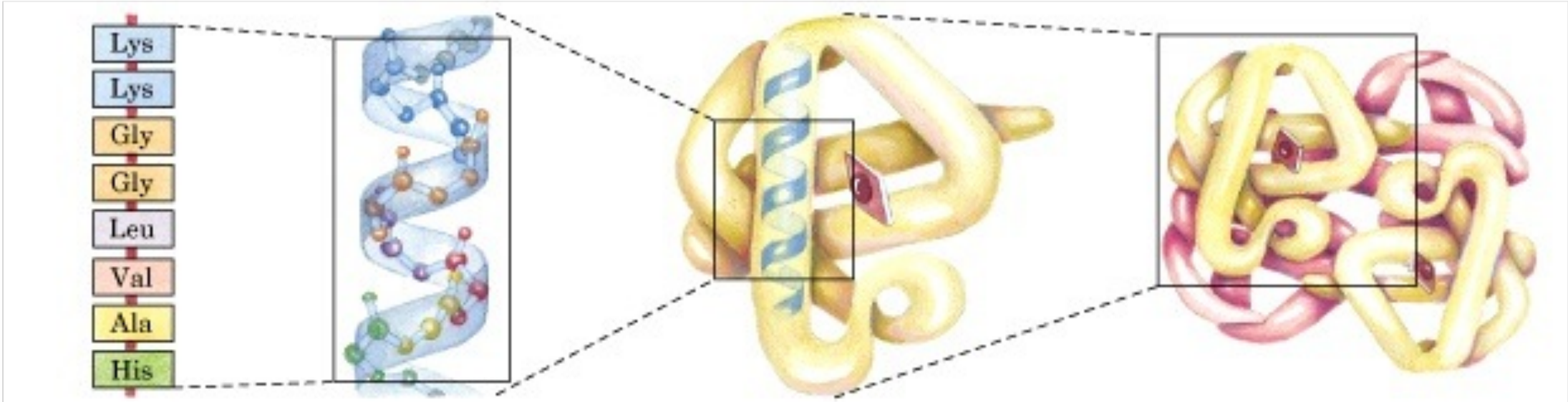
- **Overview of structural bioinformatics**
  - Motivations, goals and challenges

- **Fundamentals of protein structure**
  - Structure composition, form and forces

- **Representing, interpreting & modeling protein structure**
  - Visualizing & interpreting protein structures
  - Analyzing protein structures
  - Modeling energy as a function of structure

HIERARCHICAL STRUCTURE OF PROTEINS

Primary > Secondary > Tertiary > Quaternary



amino acid residues

Alpha helix

Polypeptide chain

Assembled subunits

Image from: http://www.ncbi.nlm.nih.gov/books/NBK21581/

RECAP: AMINO ACID NOMENCLATURE

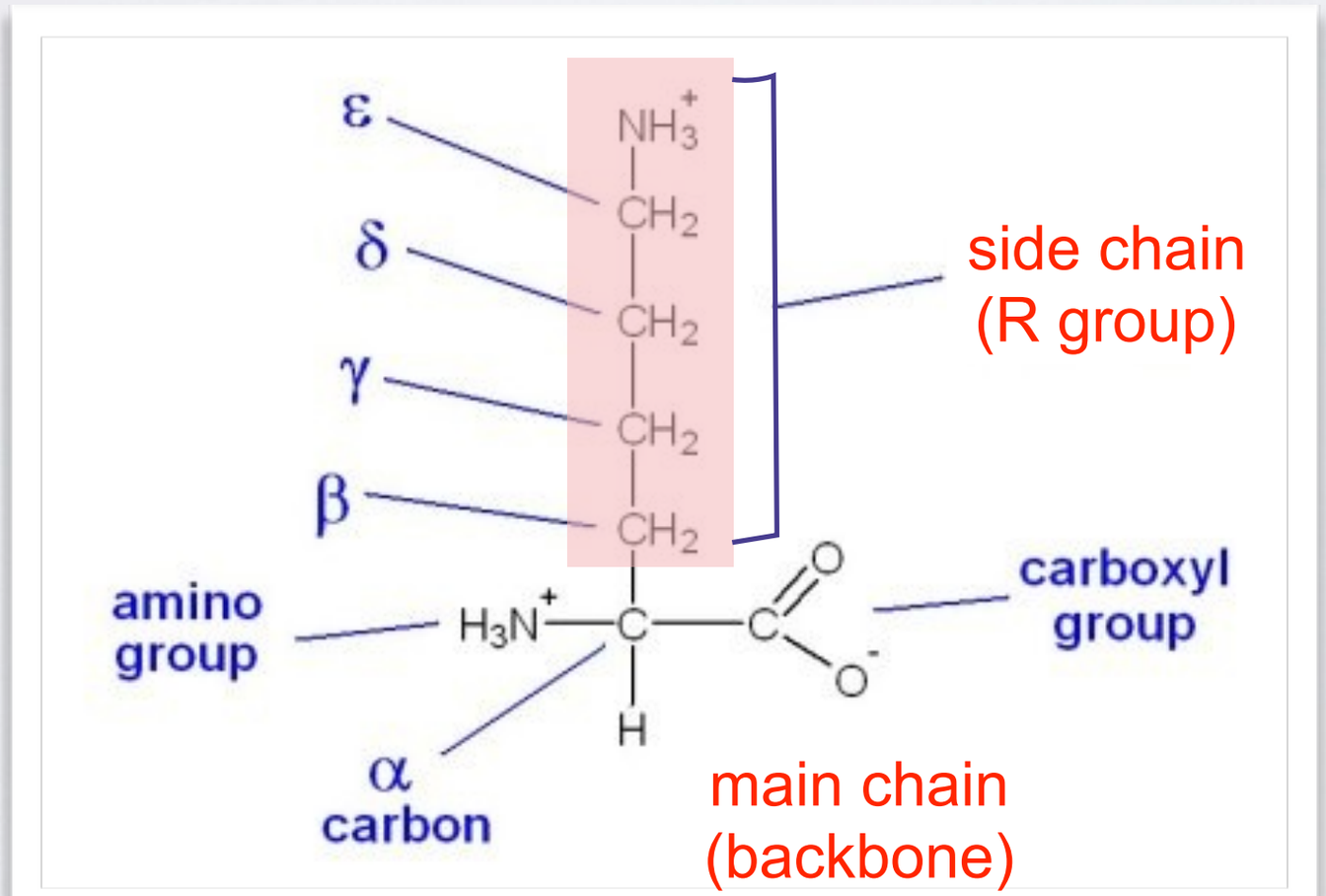
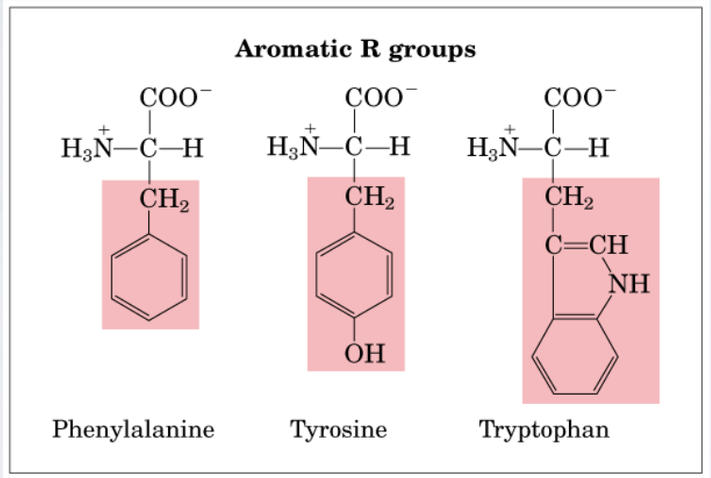
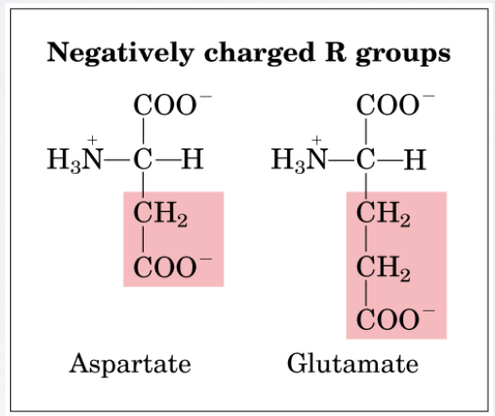
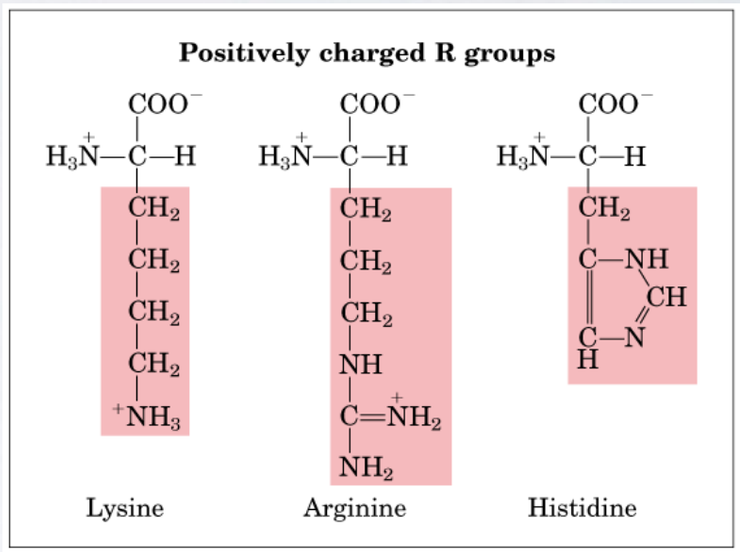


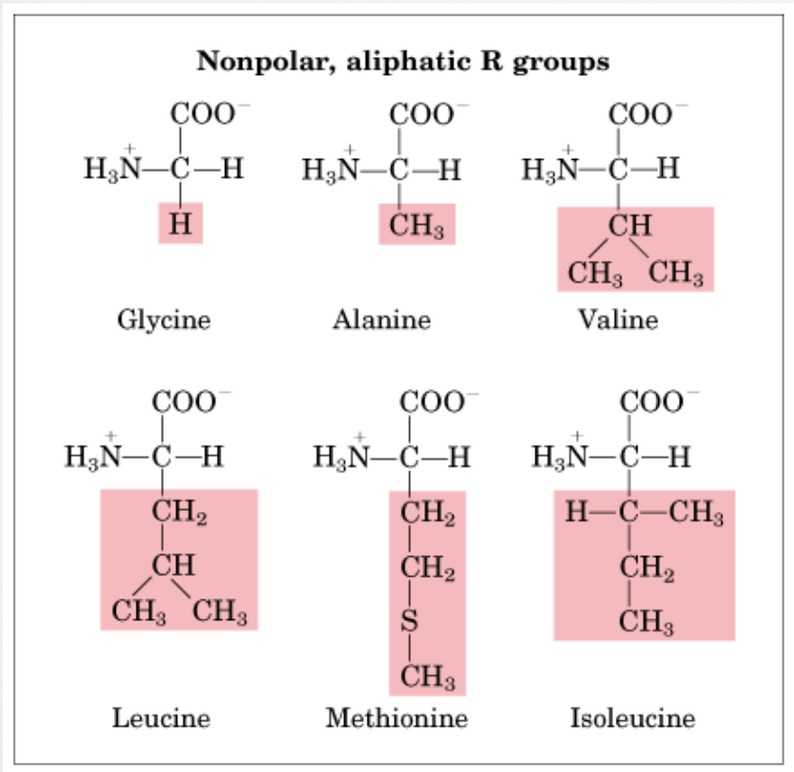
Image from: http://www.ncbi.nlm.nih.gov/books/NBK21581/

AMINO ACIDS CAN BE GROUPED BY THE **PHYSIOCHEMICAL PROPERTIES**









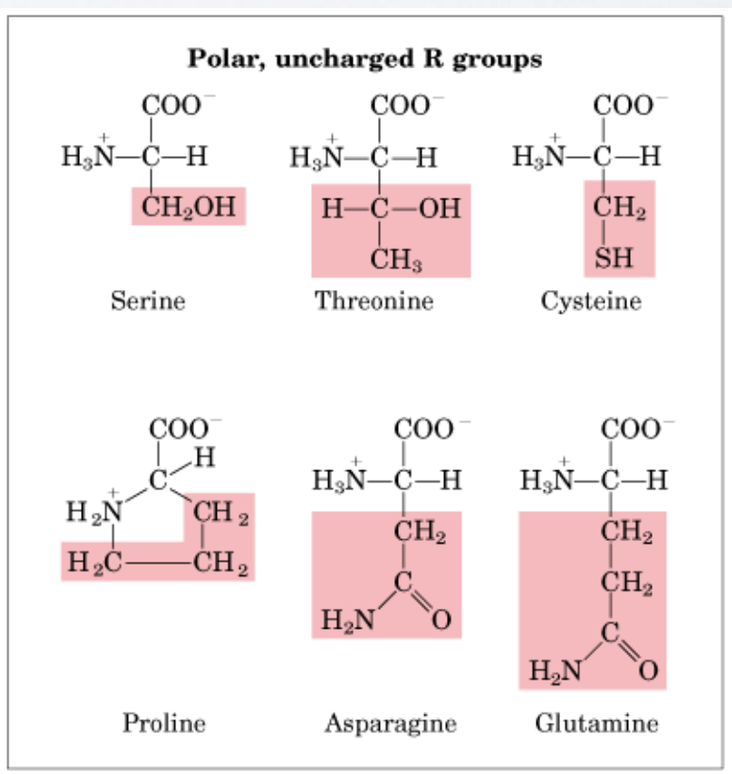
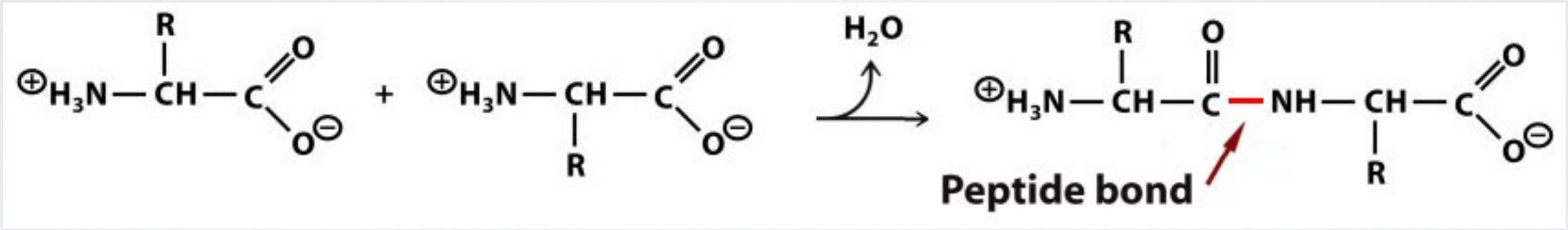


Image from: http://www.ncbi.nlm.nih.gov/books/NBK21581/

AMINO ACIDS POLYMERIZE THROUGH **PEPTIDE BOND** FORMATION



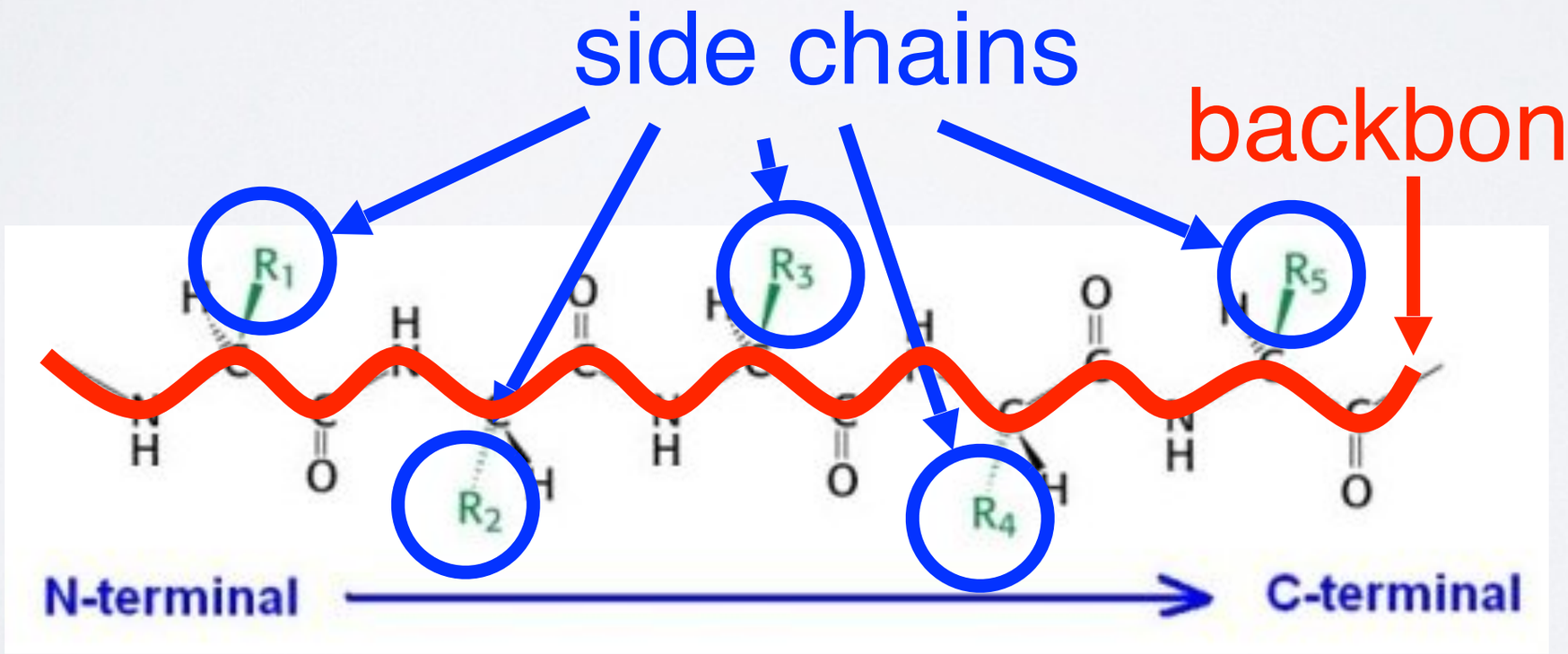
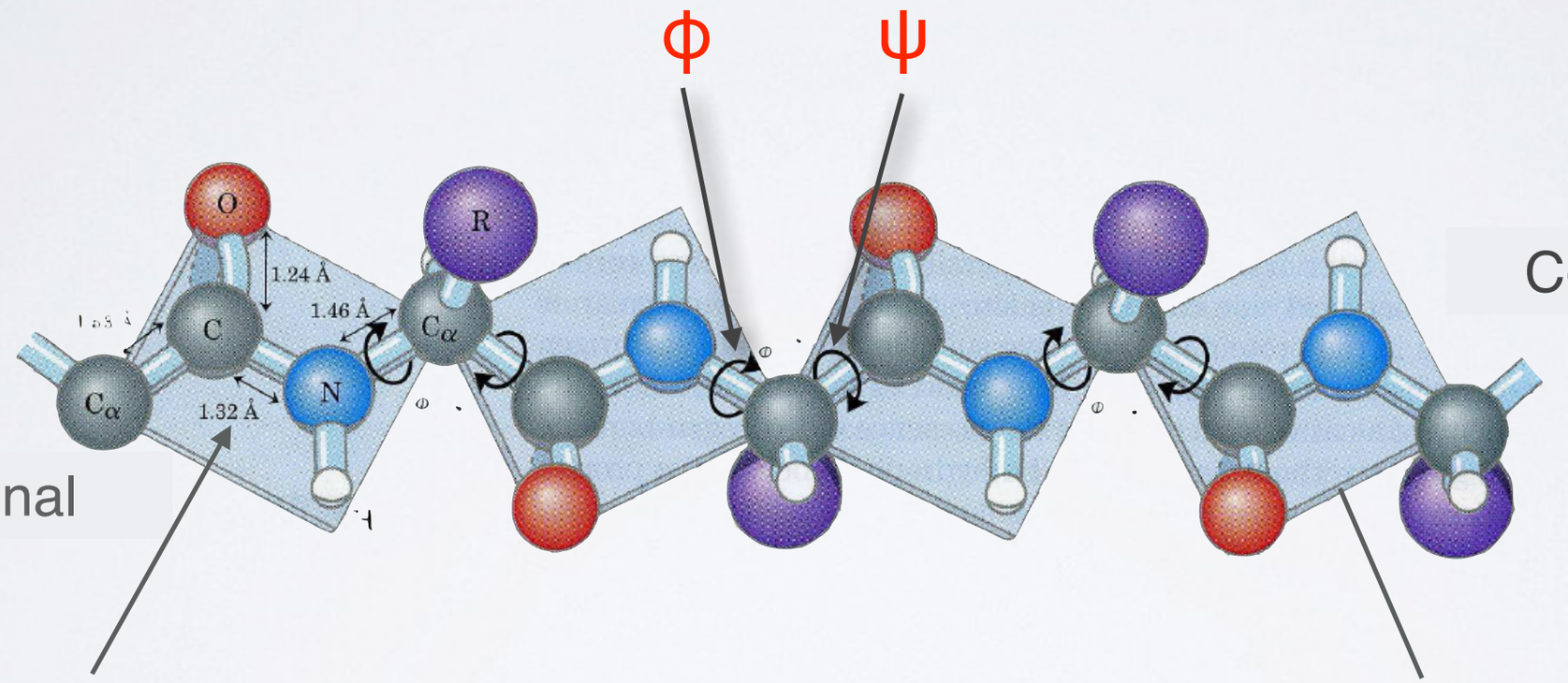


Image from: http://www.ncbi.nlm.nih.gov/books/NBK21581/

PEPTIDES CAN ADOPT DIFFERENT CONFORMATIONS BY VARYING THEIR **PHI & PSI BACKBONE TORSIONS**

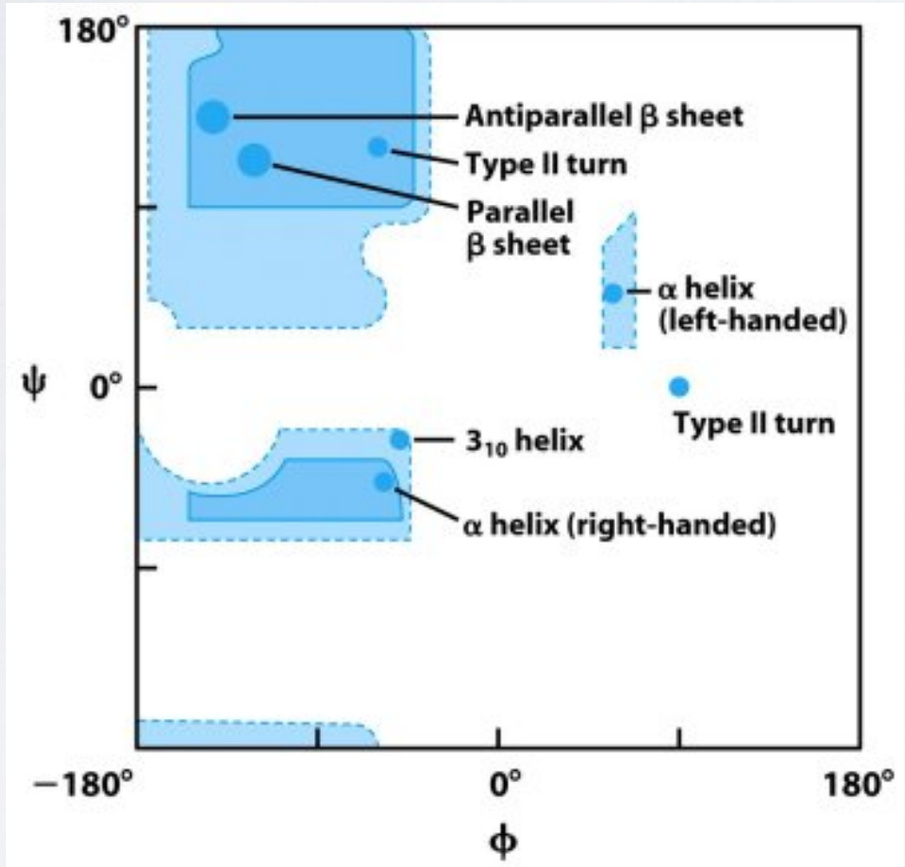


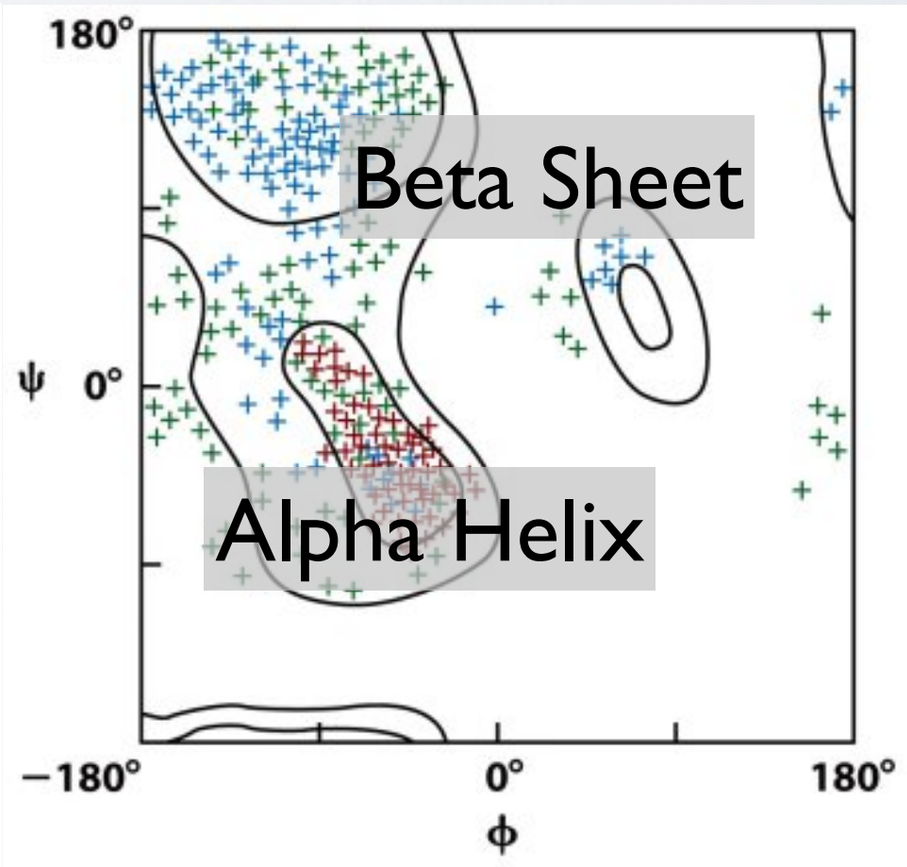
Bond angles and lengths are largely invariant

Peptide bond is planer
(Cα, C, O, N, H, Cα all lie in the same plane)

Image from: http://www.ncbi.nlm.nih.gov/books/NBK21581/

PHI vs PSI PLOTS ARE KNOWN AS **RAMACHANDRAN DIAGRAMS**

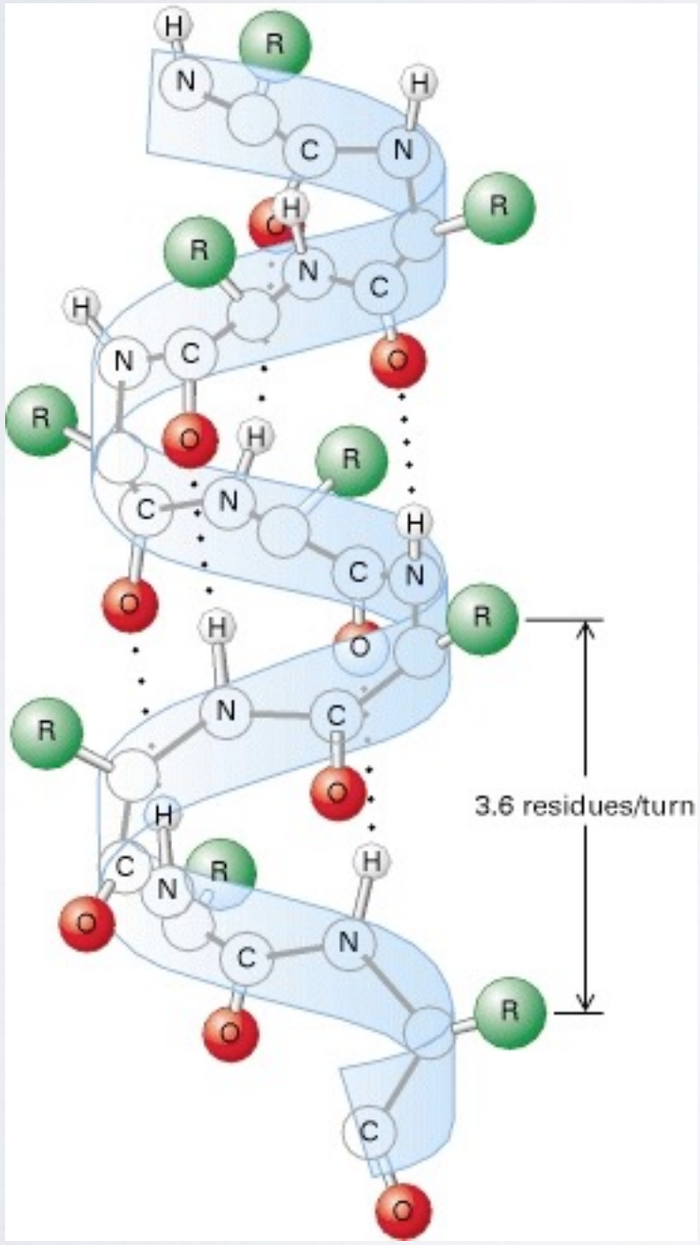




- Steric hindrance dictates torsion angle preference
- Ramachandran plot show preferred regions of ϕ and ψ dihedral angles which correspond to major forms of secondary structure

Image from: http://www.ncbi.nlm.nih.gov/books/NBK21581/

MAJOR SECONDARY STRUCTURE TYPES
ALPHA HELIX & BETA SHEET

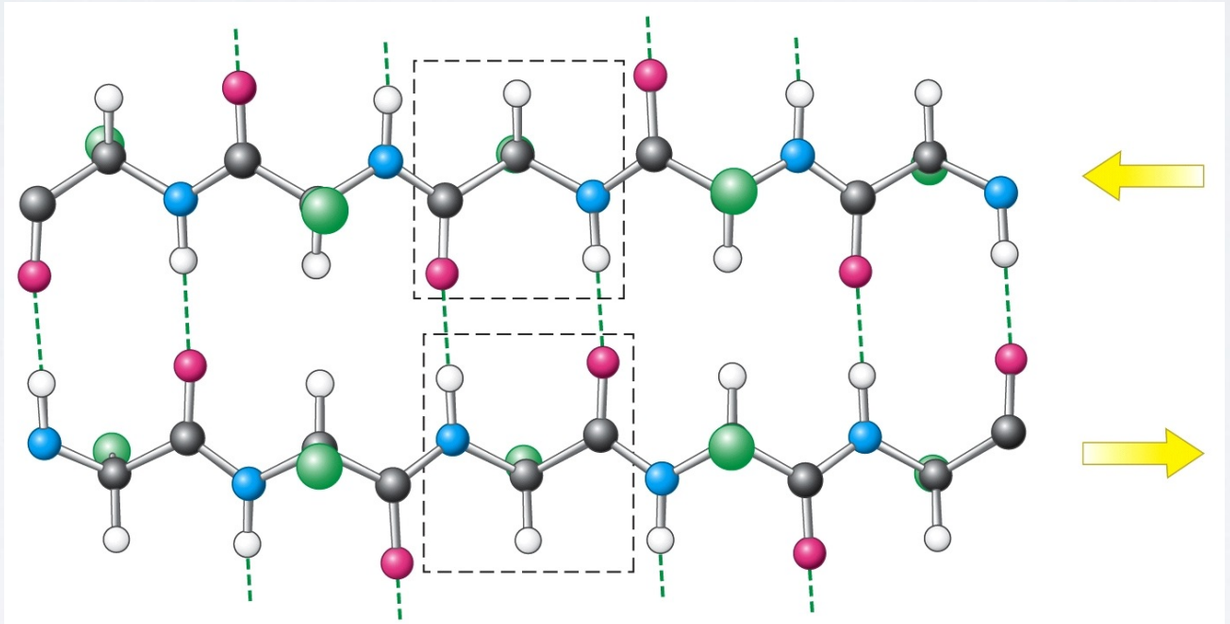


α-helix

- Most common from has <u>3.6 residues per turn</u> (number of residues in one full rotation)
- Hydrogen bonds (dashed lines) between residue <u>i and i+4</u> stabilize the structure
- The side chains (in green) protrude outward
- 3_{10}-helix and π-helix forms are less common

Image from: http://www.ncbi.nlm.nih.gov/books/NBK21581/

MAJOR SECONDARY STRUCTURE TYPES
ALPHA HELIX & **BETA SHEET**

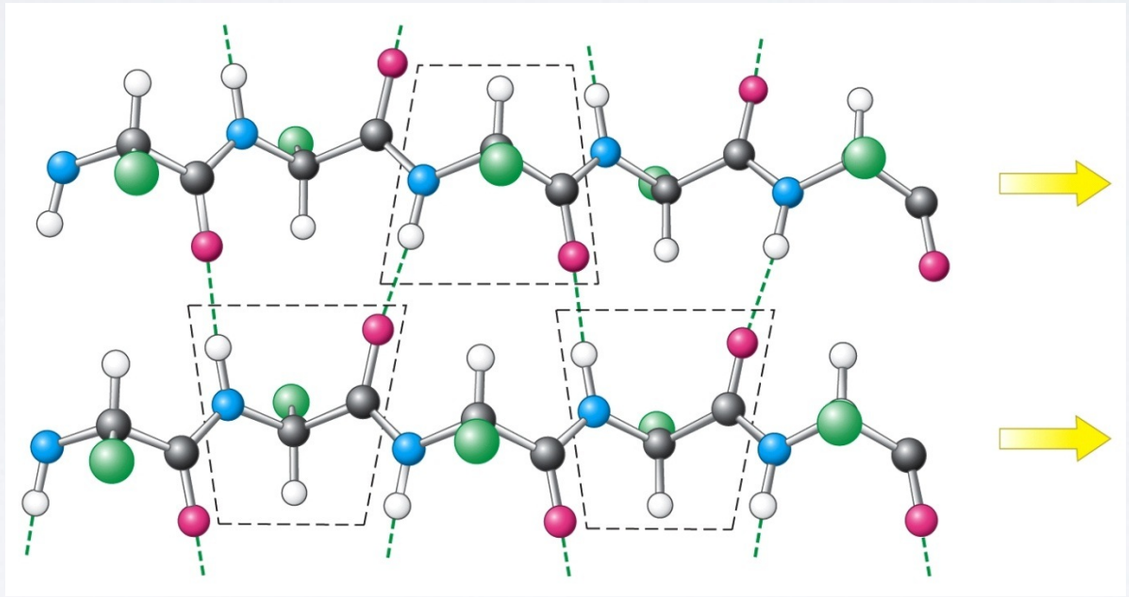


In <u>antiparallel</u> β-sheets

- Adjacent β-strands run in <u>opposite</u> directions
- Hydrogen bonds (dashed lines) between NH and CO stabilize the structure
- The side chains (in green) are above and below the sheet

Image from: http://www.ncbi.nlm.nih.gov/books/NBK21581/

MAJOR SECONDARY STRUCTURE TYPES
ALPHA HELIX & **BETA SHEET**



In <u>parallel</u> β-sheets

- Adjacent β-strands run in <u>same</u> direction
- Hydrogen bonds (dashed lines) between NH and CO stabilize the structure
- The side chains (in green) are above and below the sheet

Image from: http://www.ncbi.nlm.nih.gov/books/NBK21581/

Protein Data Bank (PDB) is the main repository for Biomolecular structure data

http://www.rcsb.org

www.rcsb.org

Home Gmail Gcal GitHub BIMM143 BGGN213 Atmosphere BIMM194 Blink News +

RCSB PDB Deposit Search Visualize Analyze Download Learn More MyPDB

RCSB PDB PROTEIN DATA BANK

153085 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

Search by PDB ID, author, macromolecule, sequence, or ligands

Go

Advanced Search | Browse by Annotations

PDB-101 wwPDB EMDataResource ndb Worldwide Protein Data Bank Foundation

Welcome

Deposit

Search

Visualize

Analyze

Download

Learn

A Structural View of Biology

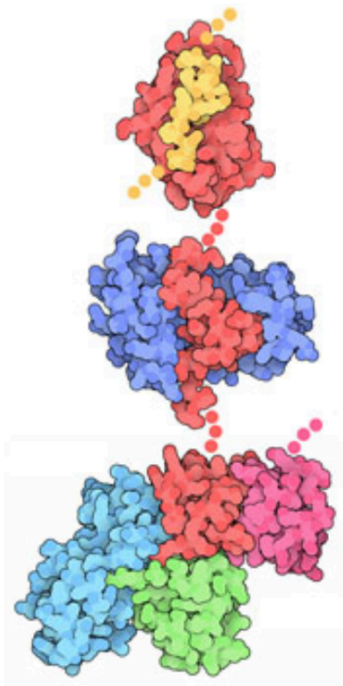
This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

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

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

High School Antibiotic Resistance Video Challenge

June Molecule of the Month



MDM2 and Cancer

Contact Us

You can search by text (e.g. "ABL kinase"), PDB code (e.g. "1iep") or sequence

http://www.rcsb.org

www.rcsb.org/structure/1iep

Home Gmail Gcal GitHub BIMM143 BGGN213 Atmosphere BIMM194 Blink News +

RCSB PDB Deposit Search Visualize Analyze Download Learn More MyPDB

RCSB PDB PROTEIN DATA BANK

153085 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

1iep

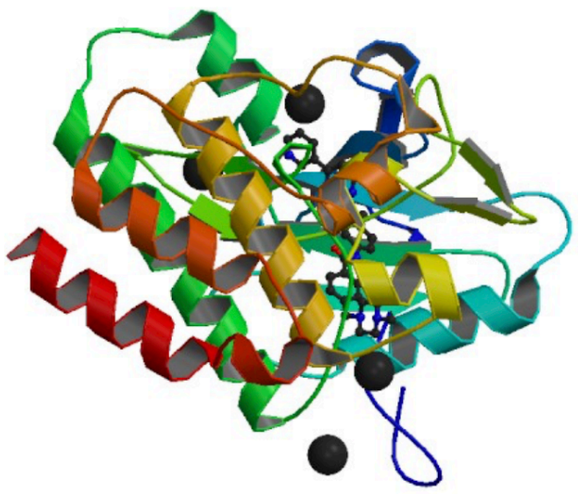
Go

Advanced Search | Browse by Annotations

PDB-101 wwPDB PROTEIN DATA BANK EMDataResource Unified Data Resource for 3DEM ndb NUCLEIC ACID DATABASE Worldwide Protein Data Bank Foundation

Structure Summary | 3D View | Annotations | Sequence | Sequence Similarity | Structure Similarity | Experiment

Biological Assembly 1



3D View: Structure | Electron Density | Ligand Interaction

Display Files | Download Files

Contact Us

1IEP

CRYSTAL STRUCTURE OF THE C-ABL KINASE DOMAIN IN COMPLEX WITH STI-571.

DOI: 10.2210/pdb1IEP/pdb

Classification: TRANSFERASE
Organism(s): Mus musculus
Expression System: Spodoptera frugiperda

Deposited: 2001-04-10 **Released:** 2001-04-18
Deposition Author(s): Nagar, B., Bornmann, W., Schindler, T., Clarkson, B., Kuriyan, J.

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 2.1 Å
R-Value Free: 0.262
R-Value Work: 0.231

wwPDB Validation 3D Report | Full Report

Metric	Percentile Ranks	Value
Rfree		0.264
Clashscore		15
Ramachandran outliers		1.3%
Sidechain outliers		2.9%
RSRZ outliers		15.7%

You can get a 3D View of and read details about the experiment and molecule

http://www.rcsb.org

www.rcsb.org/3d-view/1IEP

Home Gmail Gcal GitHub BIMM143 BGGN213 Atmosphere BIMM194 Blink News +

RCSB PDB Deposit Search Visualize Analyze Download Learn More MyPDB

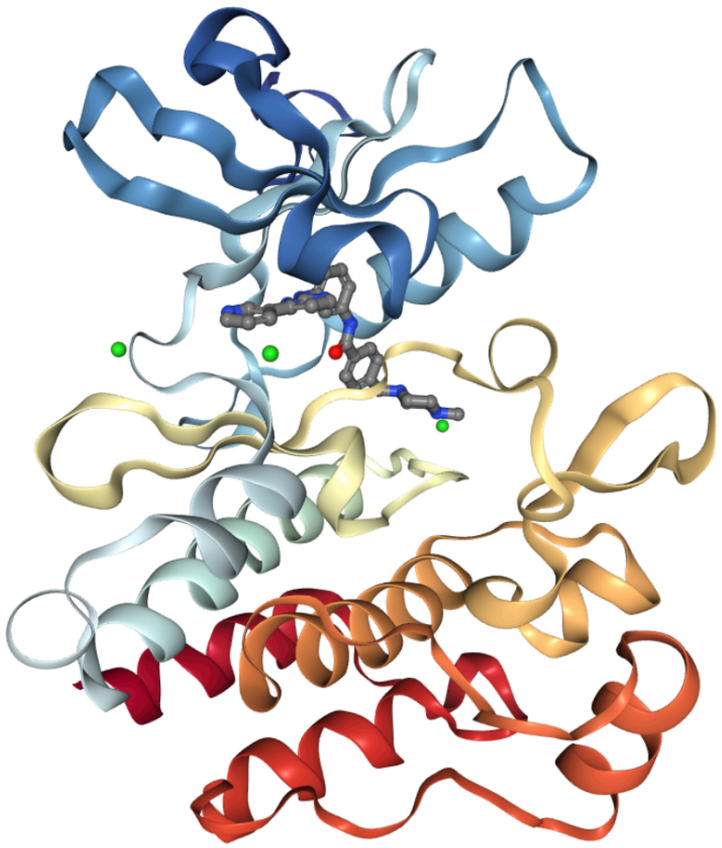
Structure Summary | 3D View | Annotations | Sequence | Sequence Similarity | Structure Similarity | Experiment

1IEP

Display Files Download Files

CRYSTAL STRUCTURE OF THE C-ABL KINASE DOMAIN IN COMPLEX WITH STI-571.

Note: Use your mouse to drag, rotate, and zoom in and out of the structure. Mouse-over to identify atoms and bonds. Mouse controls documentation.



Structure View | Electron Density Maps | Ligand View

Structure View Documentation

Assembly: Bioassembly 1

Model: Model 1

Symmetry: None

Style: Cartoon

Color: Rainbow

Ligand: Ball & Stick

Quality: Automatic

Water ☐ Ions ☑

Hydrogens ☑ Clashes ☐

Default Structure View

Contact Us

You can display or download PDB format files for a particular entry

http://www.rcsb.org

www.rcsb.org/3d-view/1IEP

Home Gmail Gcal GitHub BIMM143 BGGN213 Atmosphere BIMM194 Blink News +

RCSB PDB Deposit Search Visualize Analyze Download Learn More MyPDB

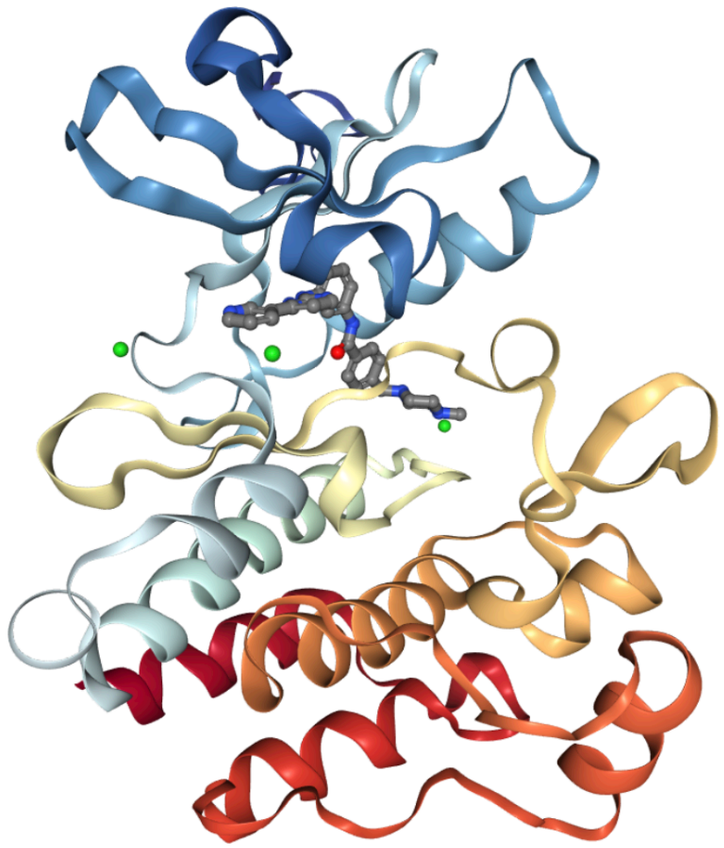
Structure Summary | 3D View | Annotations | Sequence | Sequence Similarity | Structure Similarity | Experiment

1IEP

CRYSTAL STRUCTURE OF THE C-ABL KINASE DOMAIN IN COMPLEX WITH STI-571.

Display Files | Download Files

Note: Use your mouse to drag, rotate, and zoom in and out of the structure. Mouse-over to identify atoms and bonds. Mouse controls documentation.



Structure View | Electron Density Maps | Ligand View

Structure View Documentation

Assembly: Bioassembly 1

Model: Model 1

Symmetry: None

Style: Cartoon

Color: Rainbow

Ligand: Ball & Stick

Quality: Automatic

Water ☐ Ions ☑

Hydrogens ☑ Clashes ☐

Default Structure View

Contact Us

Side-Note: PDB File Format

- PDB files contains atomic coordinates and associated information.

Amino Acid

Sequence/Residue Number

Element

Chain

Coordinates

X **Y** **Z**

(etc.)

```
ATOM      1  N   MET A   1      19.353  41.547  -3.887  ...
ATOM      2  CA  MET A   1      20.513  40.939  -4.592  ...
ATOM      3  C   MET A   1      20.150  39.658  -5.355  ...
ATOM      4  O   MET A   1      19.053  39.551  -5.903  ...
ATOM      5  CB  MET A   1      21.642  40.678  -3.592  ...
ATOM      6  CG  MET A   1      21.233  39.903  -2.360  ...
ATOM      7  SD  MET A   1      22.533  39.928  -1.113  ...
ATOM      8  CE  MET A   1      23.771  38.881  -1.885  ...
ATOM      9  N   ASP A   2      21.068  38.694  -5.390  ...
ATOM     10  CA  ASP A   2      20.856  37.440  -6.117  ...
ATOM     11  C   ASP A   2      20.124  36.371  -5.299  ...
ATOM     12  O   ASP A   2      20.680  35.818  -4.351  ...
```

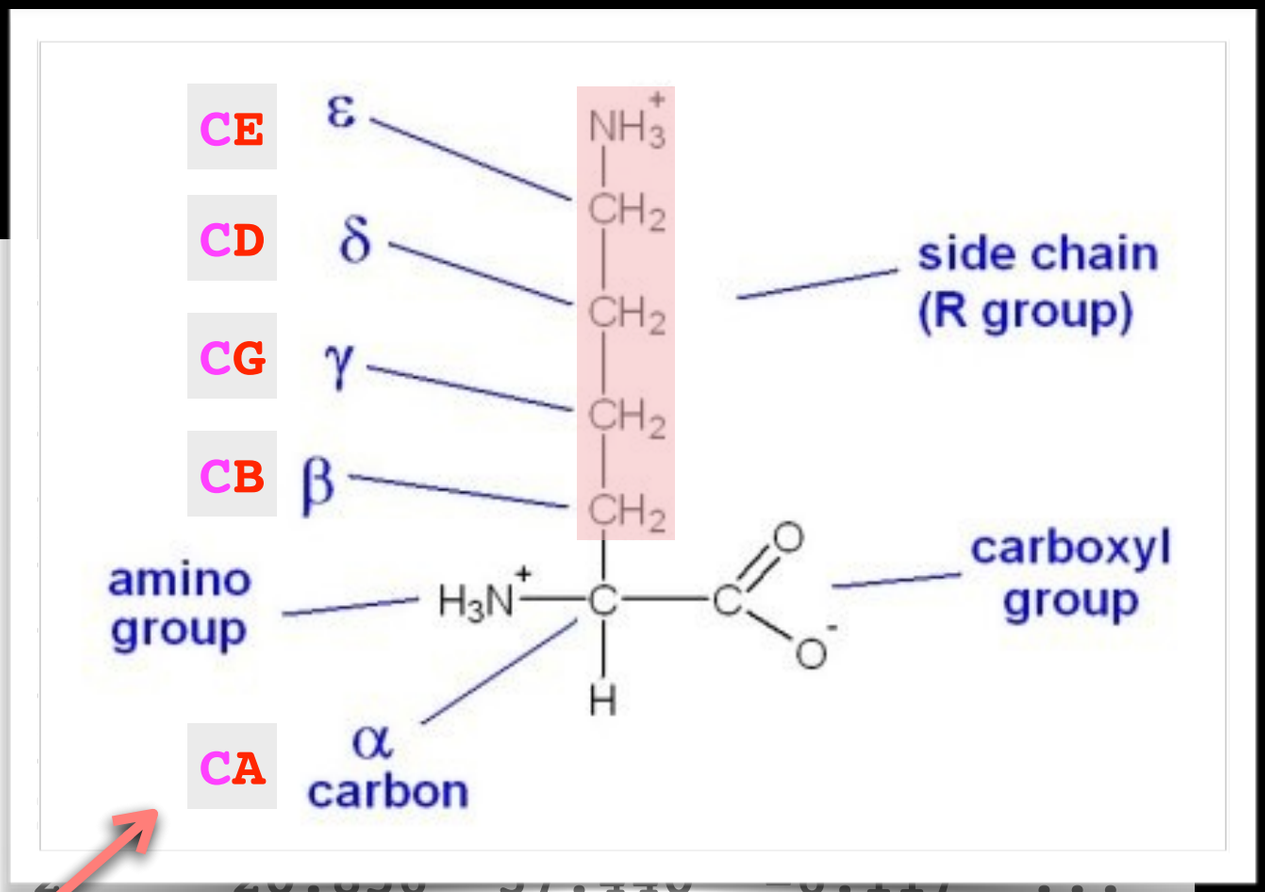
Element position within amino acid

Side-Note: PDB File Format

- PDB files contains atomic coordinates and associated information.

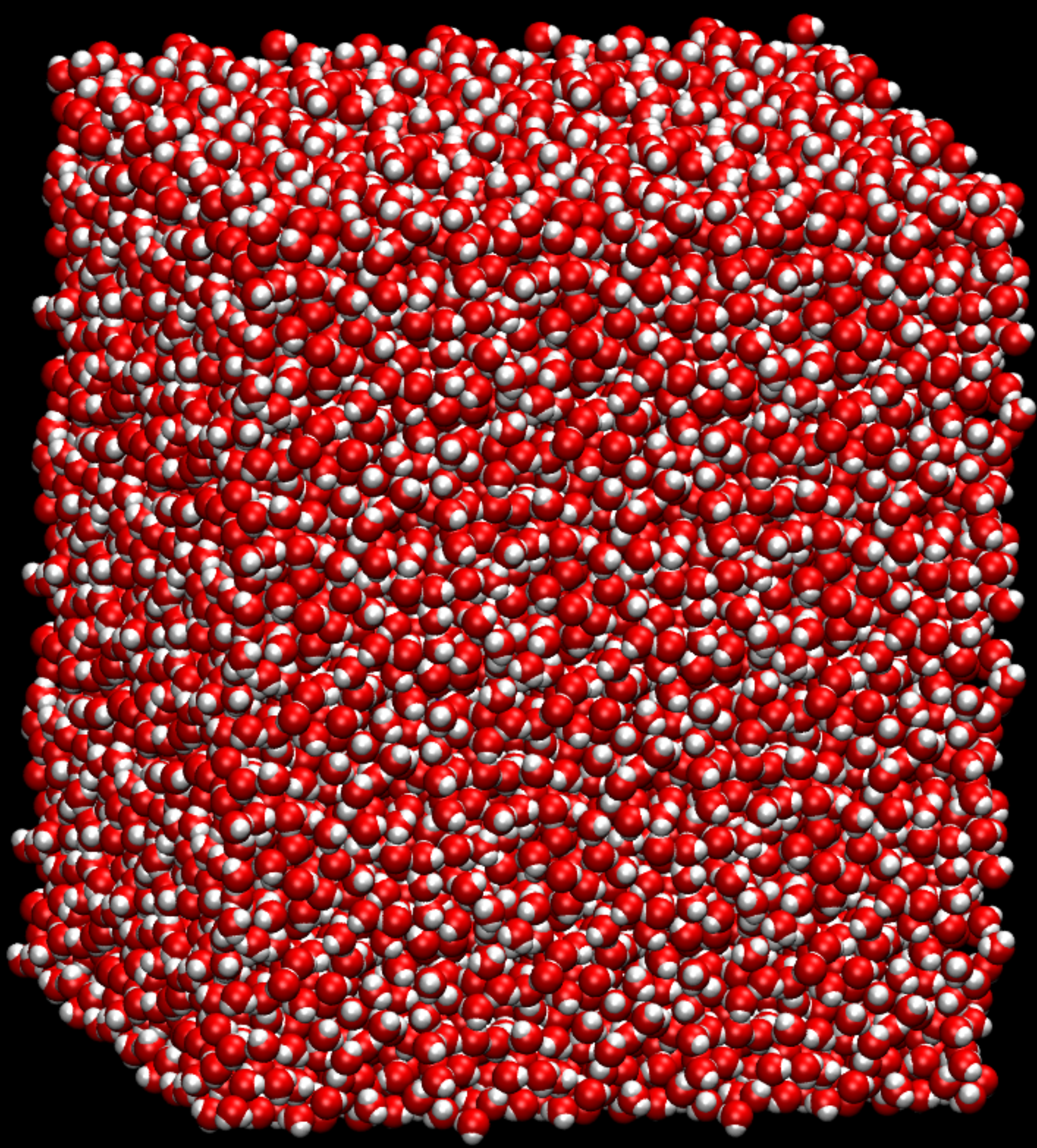
Element

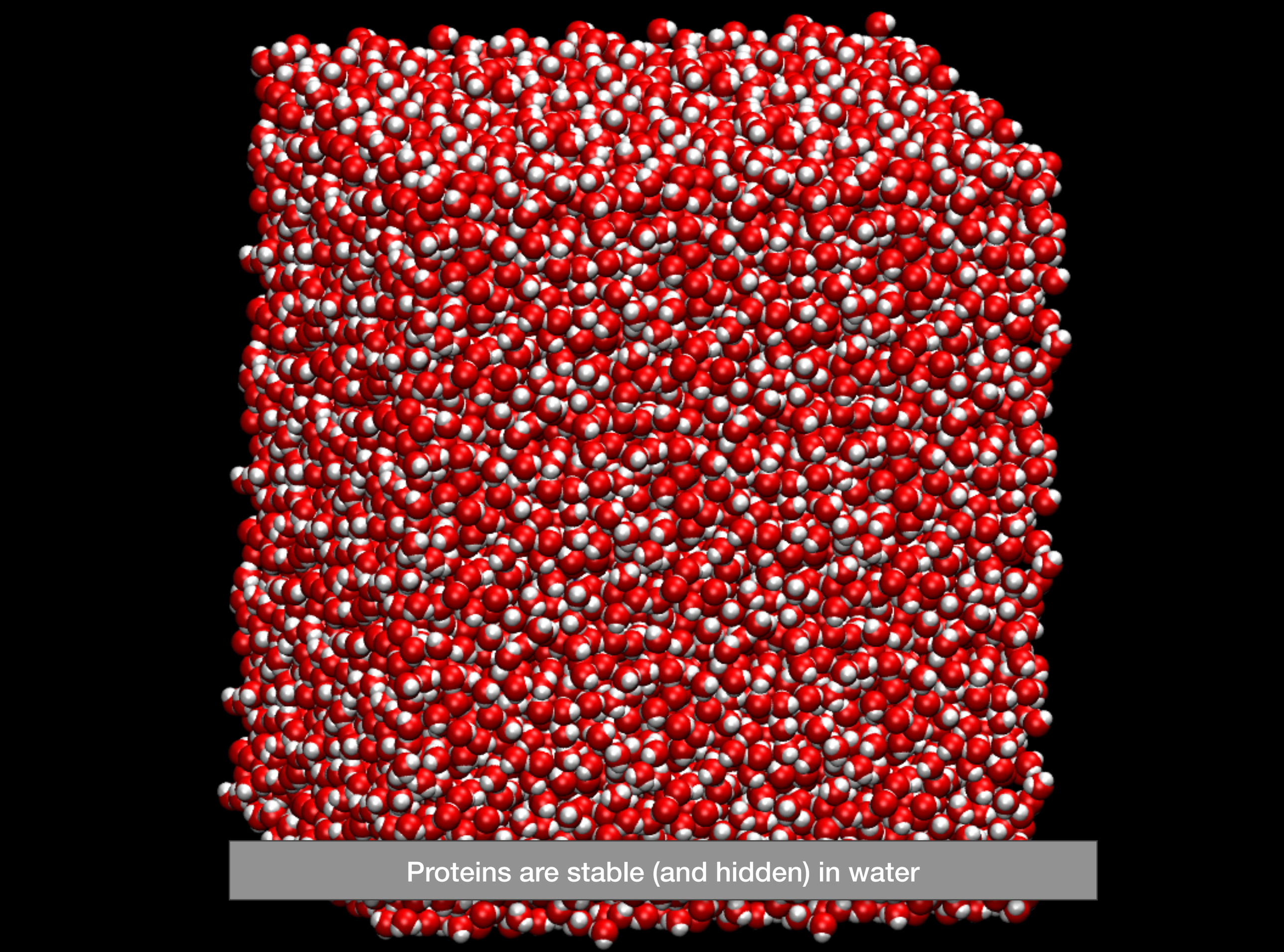
```
ATOM      1  N    MET A
ATOM      2  CA   MET A
ATOM      3  C    MET A
ATOM      4  O    MET A
ATOM      5  CB   MET A
ATOM      6  CG   MET A
ATOM      7  SD   MET A
ATOM      8  CE   MET A
ATOM      9  N    ASP A
ATOM     10  CA   ASP A
ATOM     11  C    ASP A       2      20.124  36.371  -5.299  ...
ATOM     12  O    ASP A       2      20.680  35.818  -4.351  ...
```

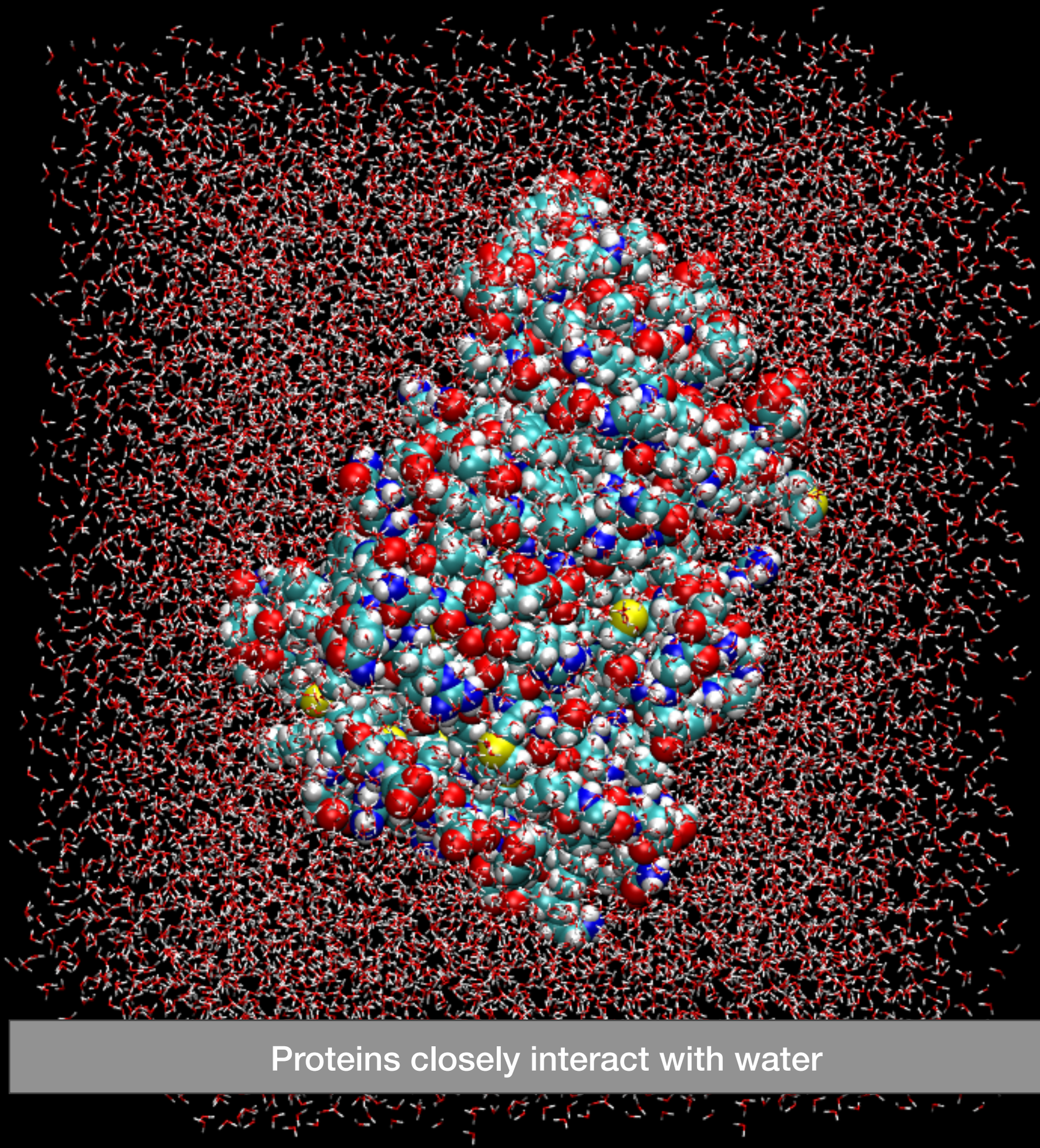


Element position within amino acid

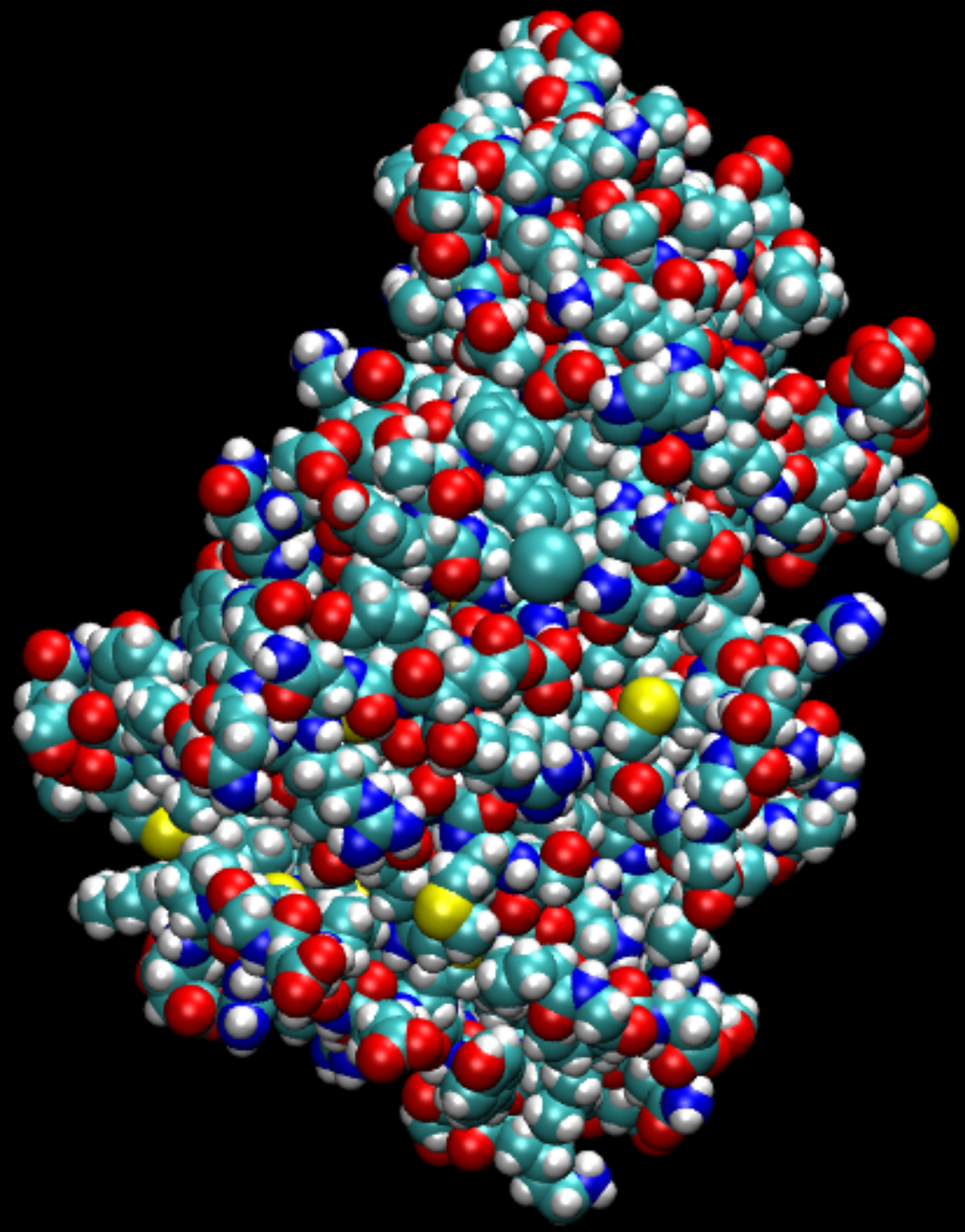
What Does a Protein Look like?



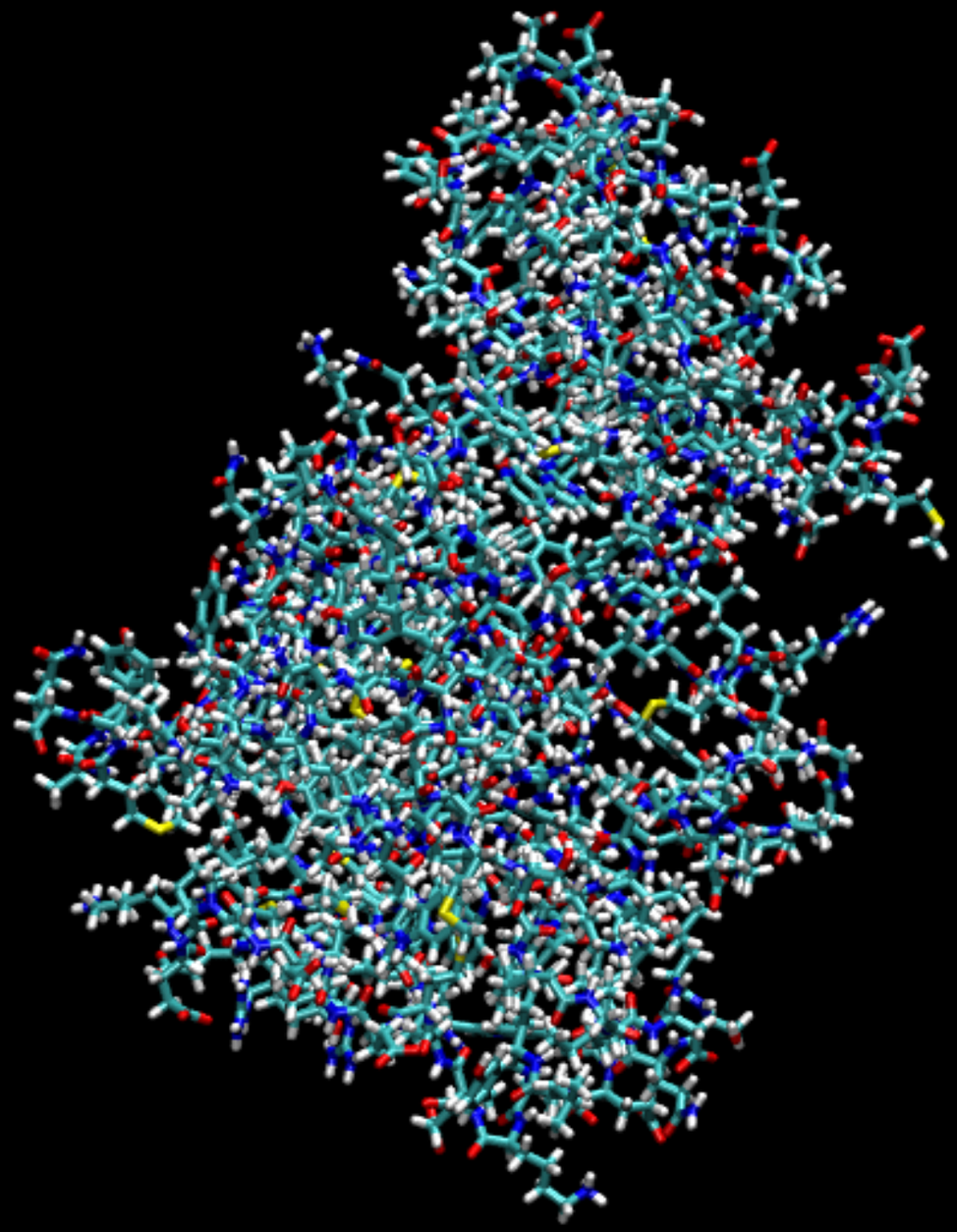

Proteins are stable (and hidden) in water



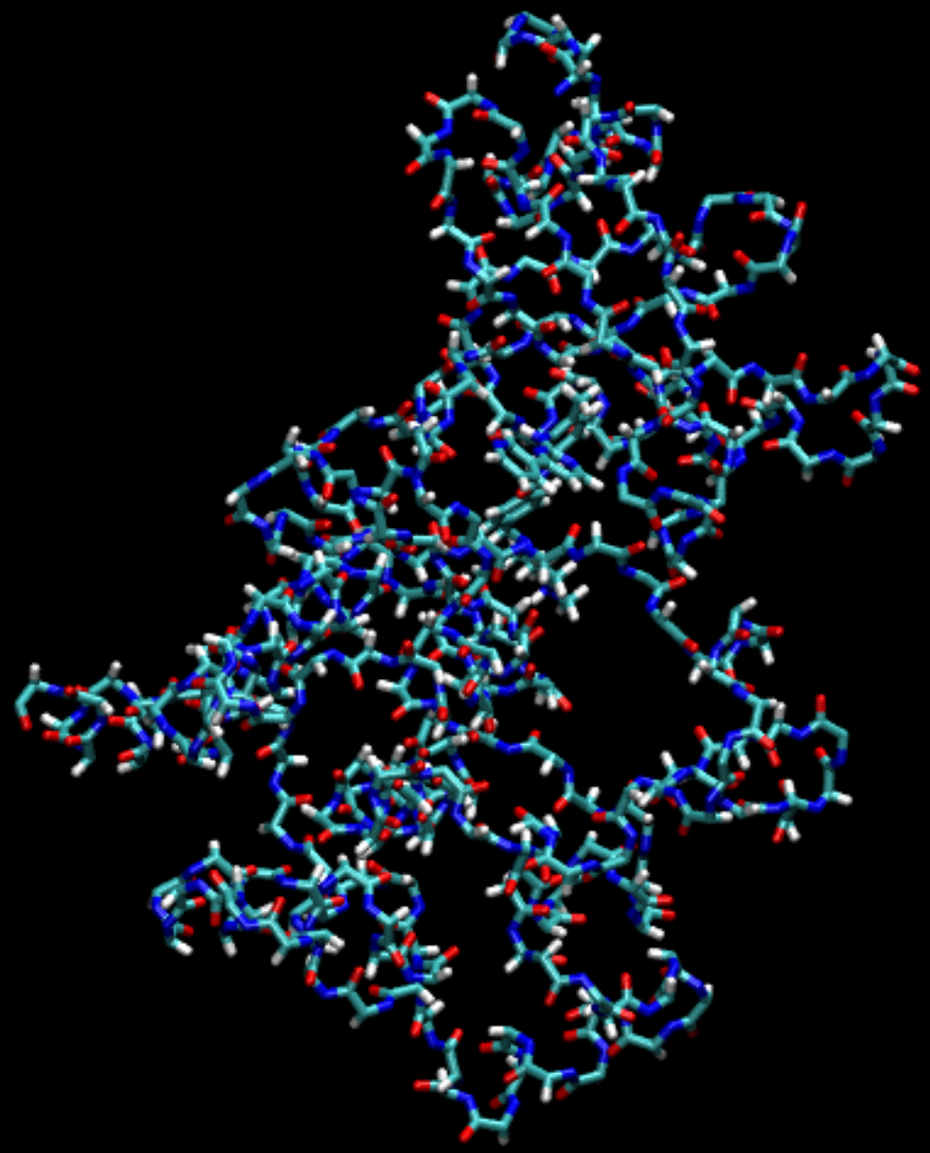
Proteins closely interact with water



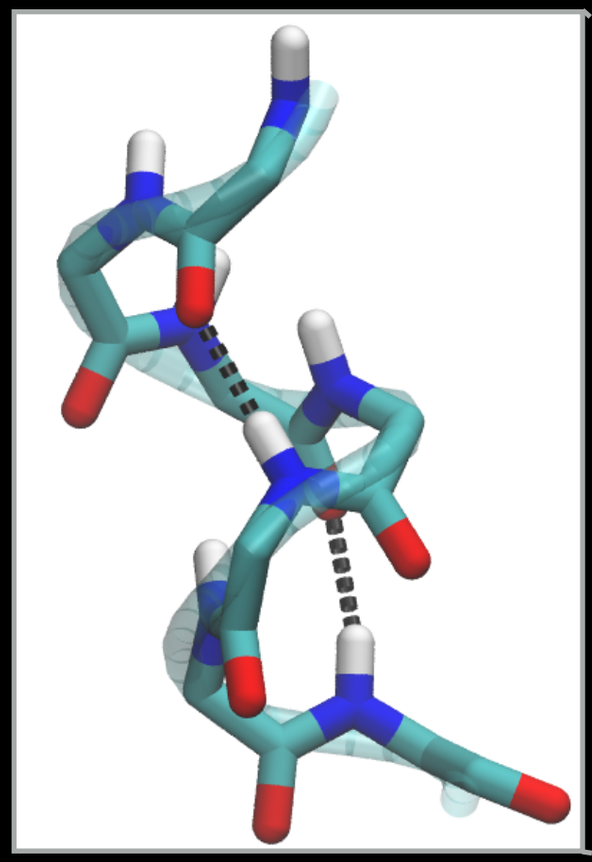
Proteins are close packed solid but flexible objects (globular)

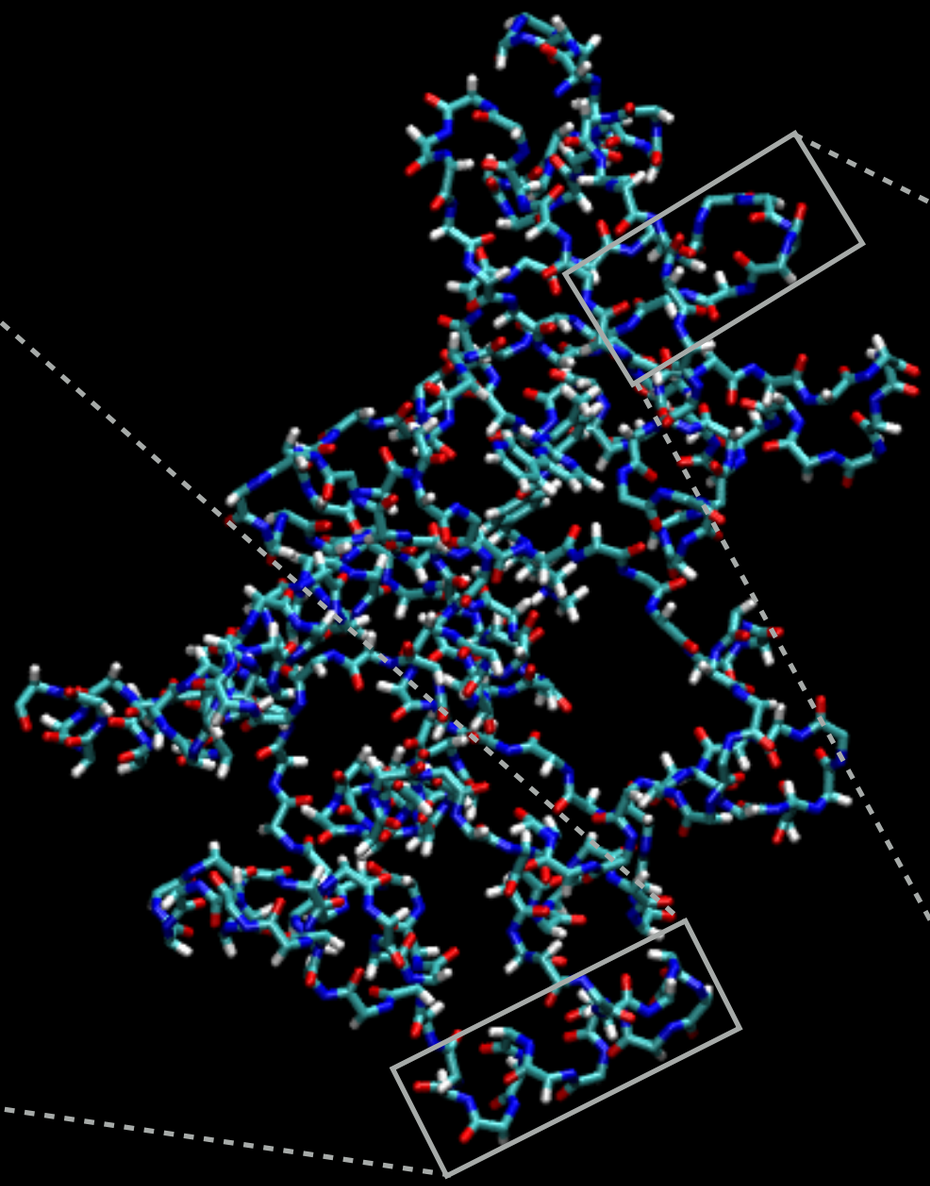


Due to their large size and complexity it is often hard to see whats important in the structure

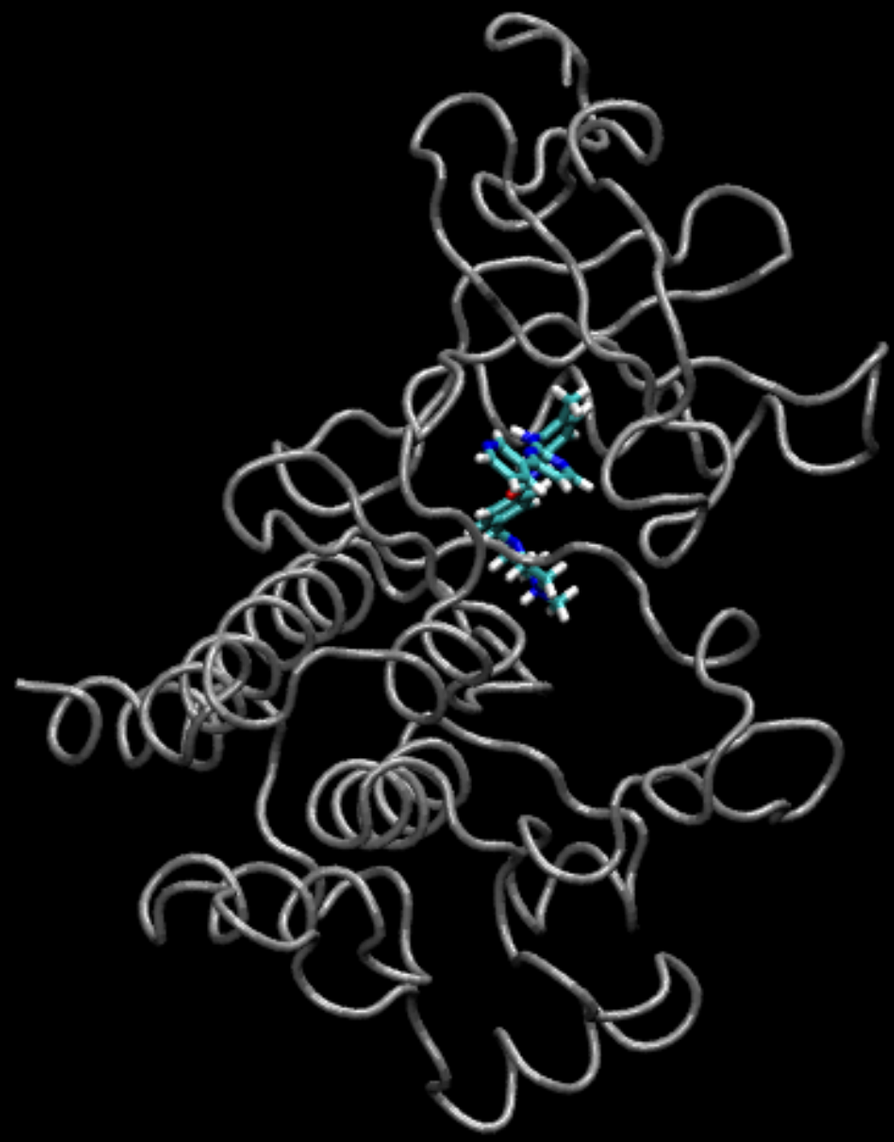


Backbone or main-chain representation can help trace chain topology





Backbone or main-chain representation can help trace chain topology & reveal secondary structure

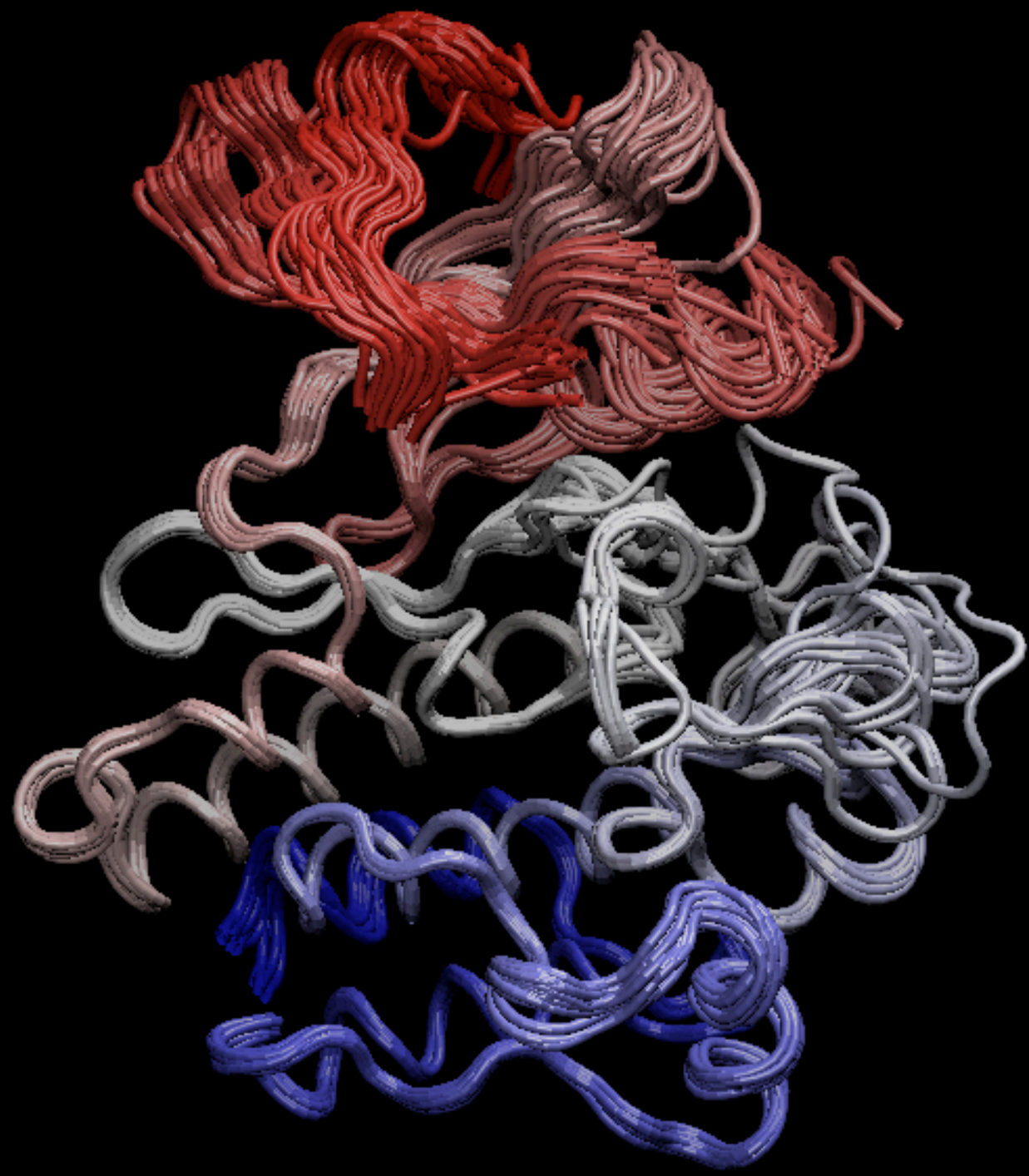


Tube or trace representation is one of the simplest views

Tube with added colors to highlight secondary structure

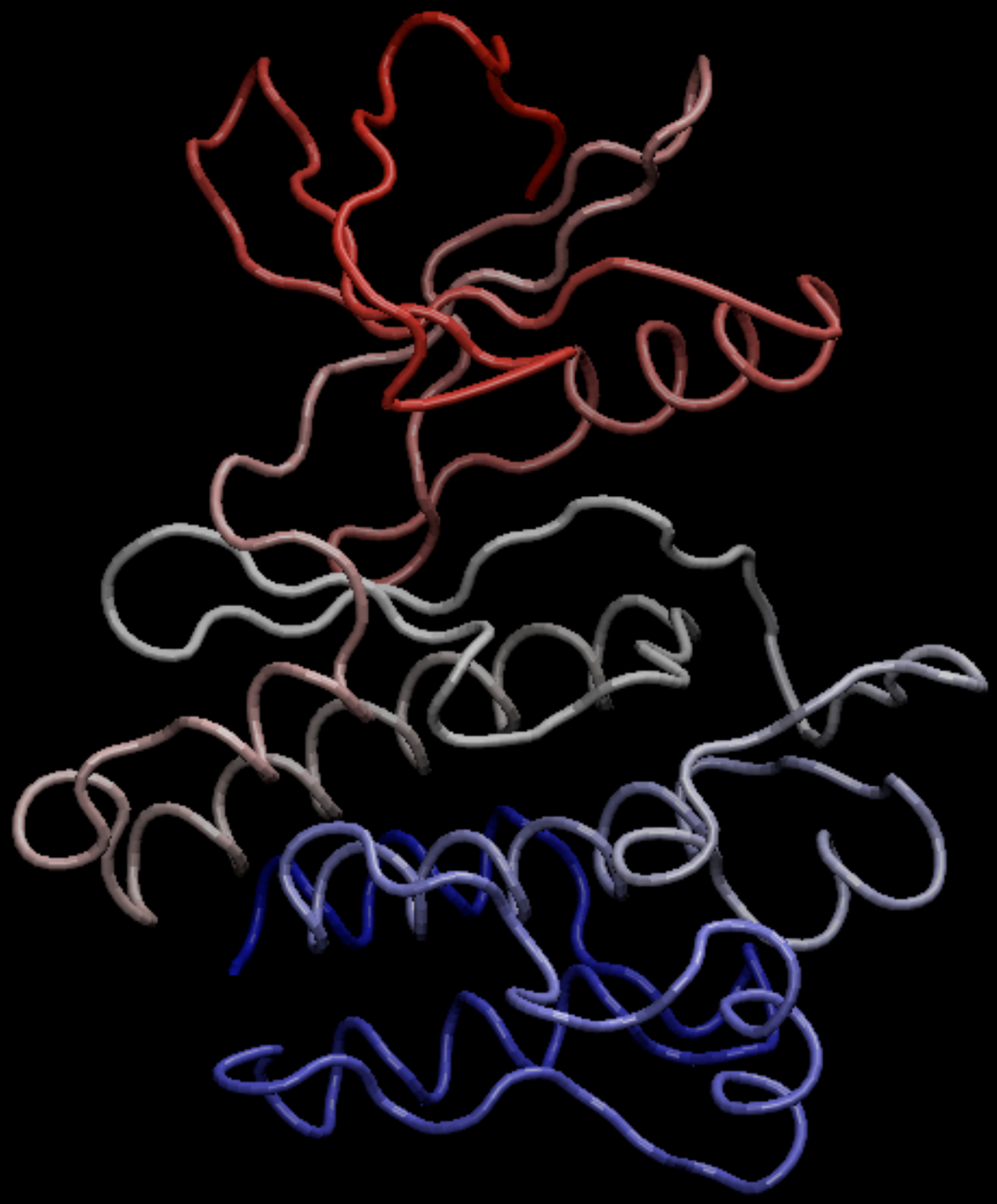
Simplified "cartoon" secondary structure representations are commonly used to communicate structural details

Viewing in 3D is often essential for interpretation.
Now we can clearly see 2º and 3º structure - the coiled chain of connected secondary structures

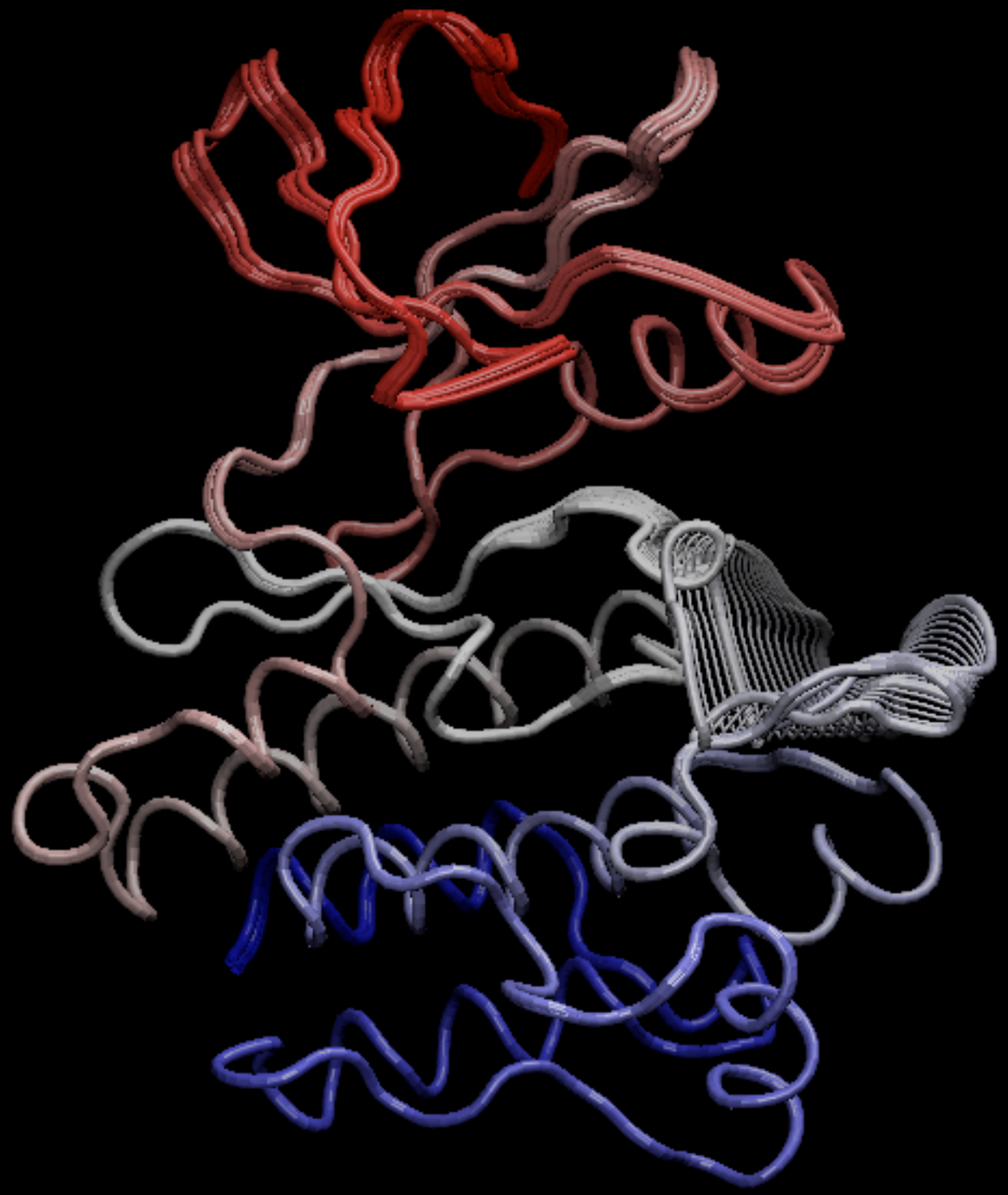


Viewing multiple superposed structures solved under different conditions can highlight **flexible regions**

Viewing multiple superposed structures solved under different conditions can highlight **distinct conformations**



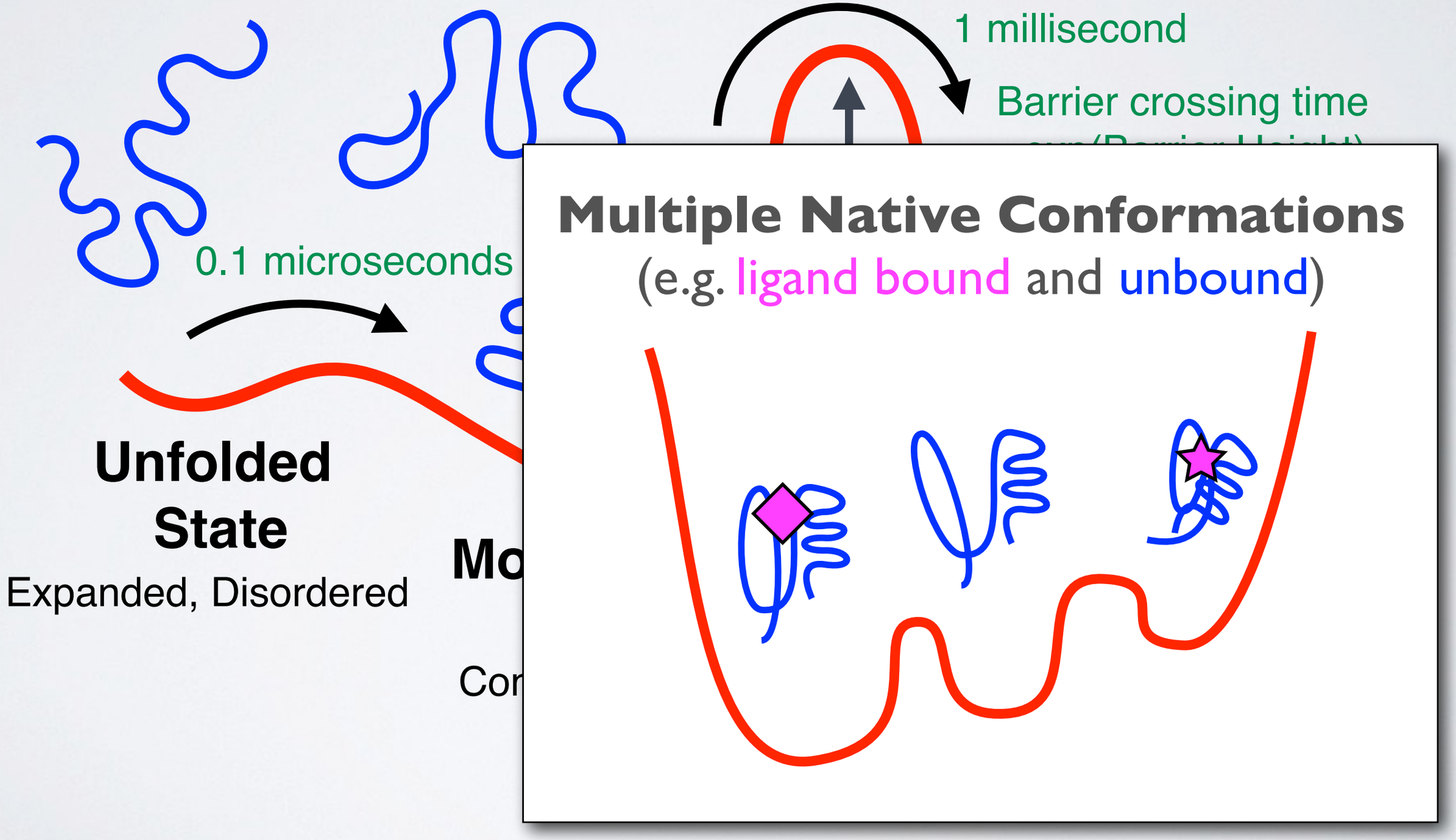
Analyzing these multiple structures can reveal **functional motions**
- i.e. displacements that are essential for regulating function



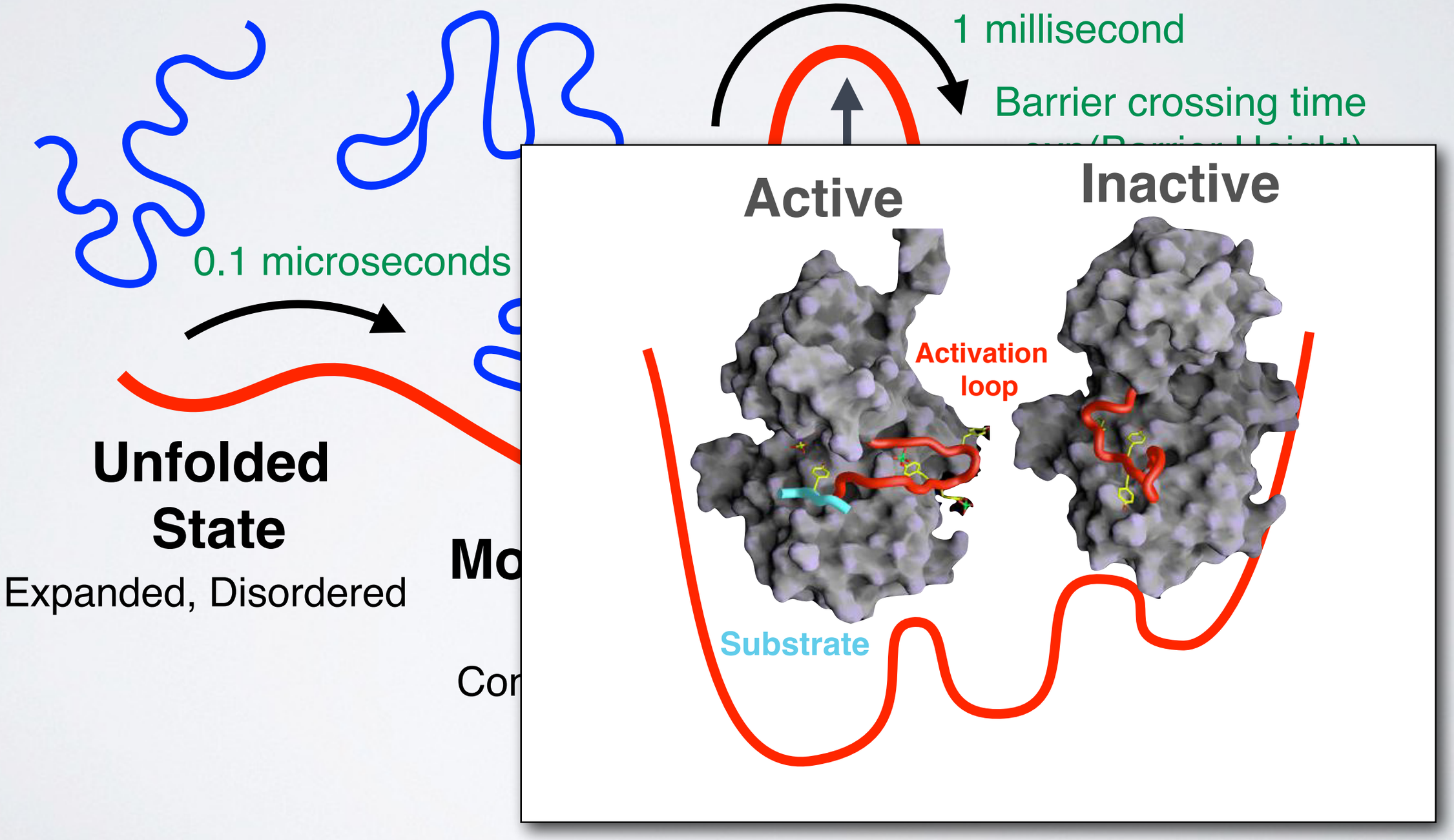
"Activation loop"

Analyzing these multiple structures can reveal **functional motions**
- i.e. displacements that are essential for regulating function

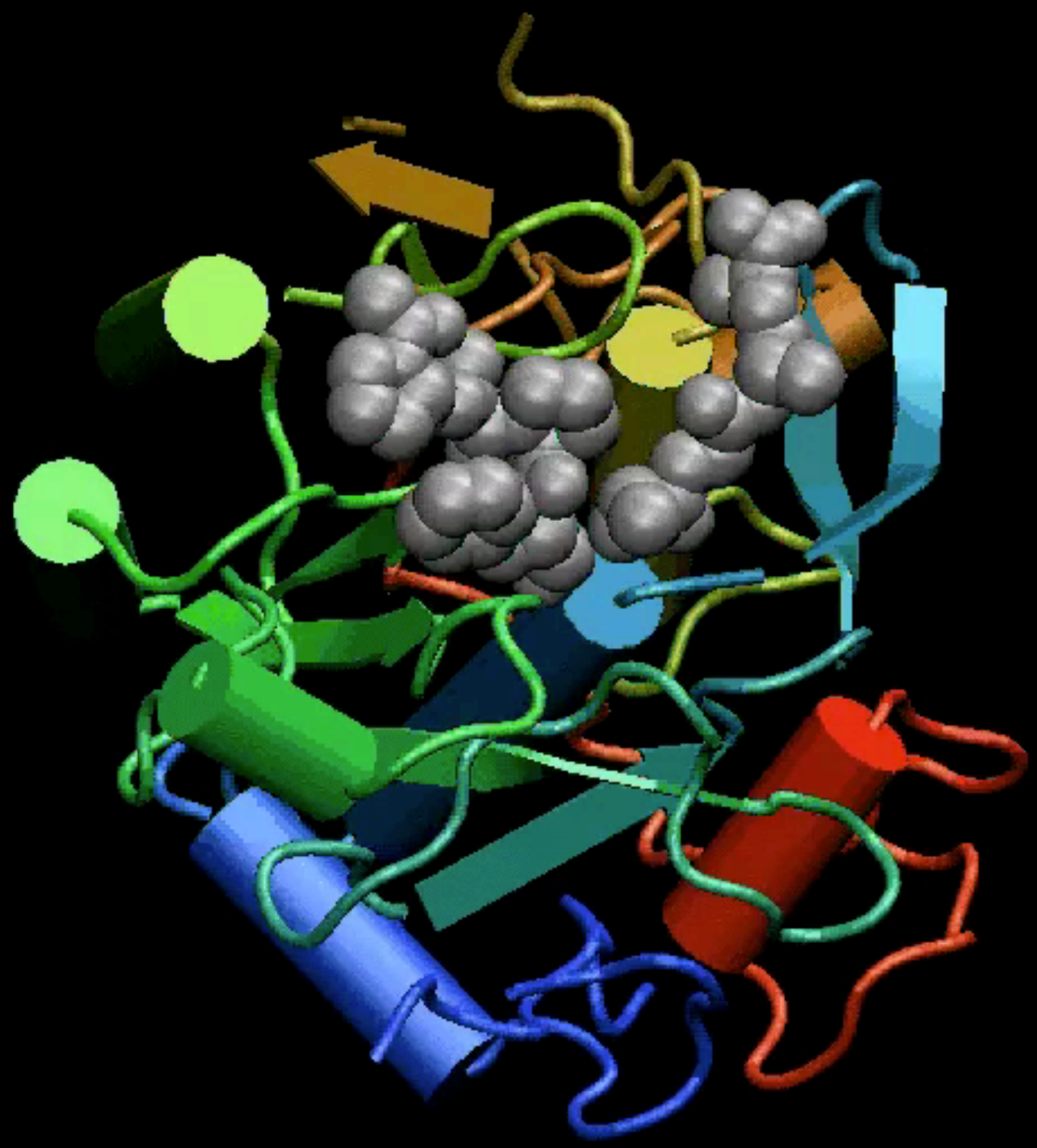
KEY CONCEPT: **ENERGY LANDSCAPE**



KEY CONCEPT: **ENERGY LANDSCAPE**



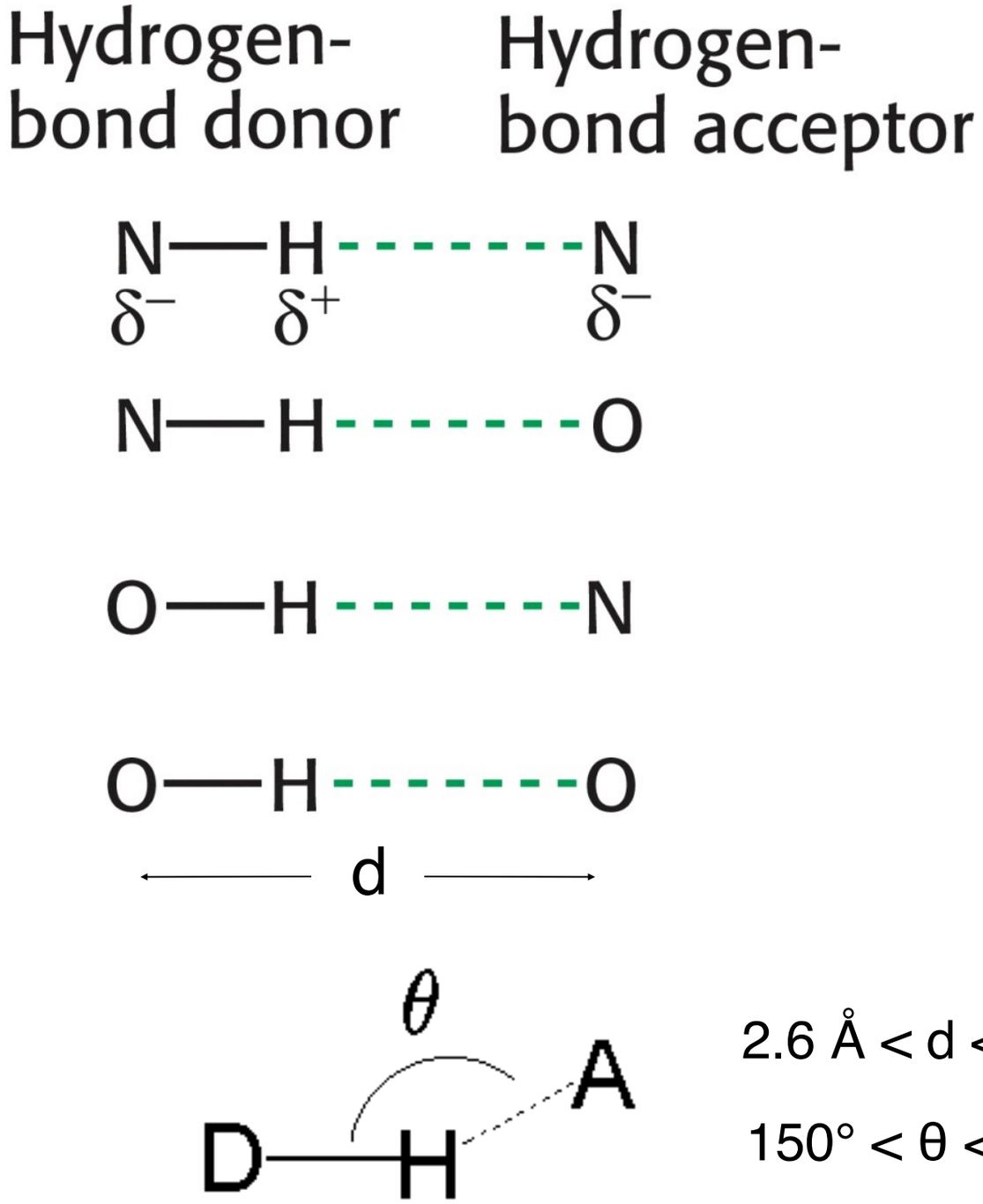
Normal **M**ode **A**nalysis (NMA) models the protein as a network of elastic strings



NMA is a bioinformatics method to predict the intrinsic dynamics of biomolecules

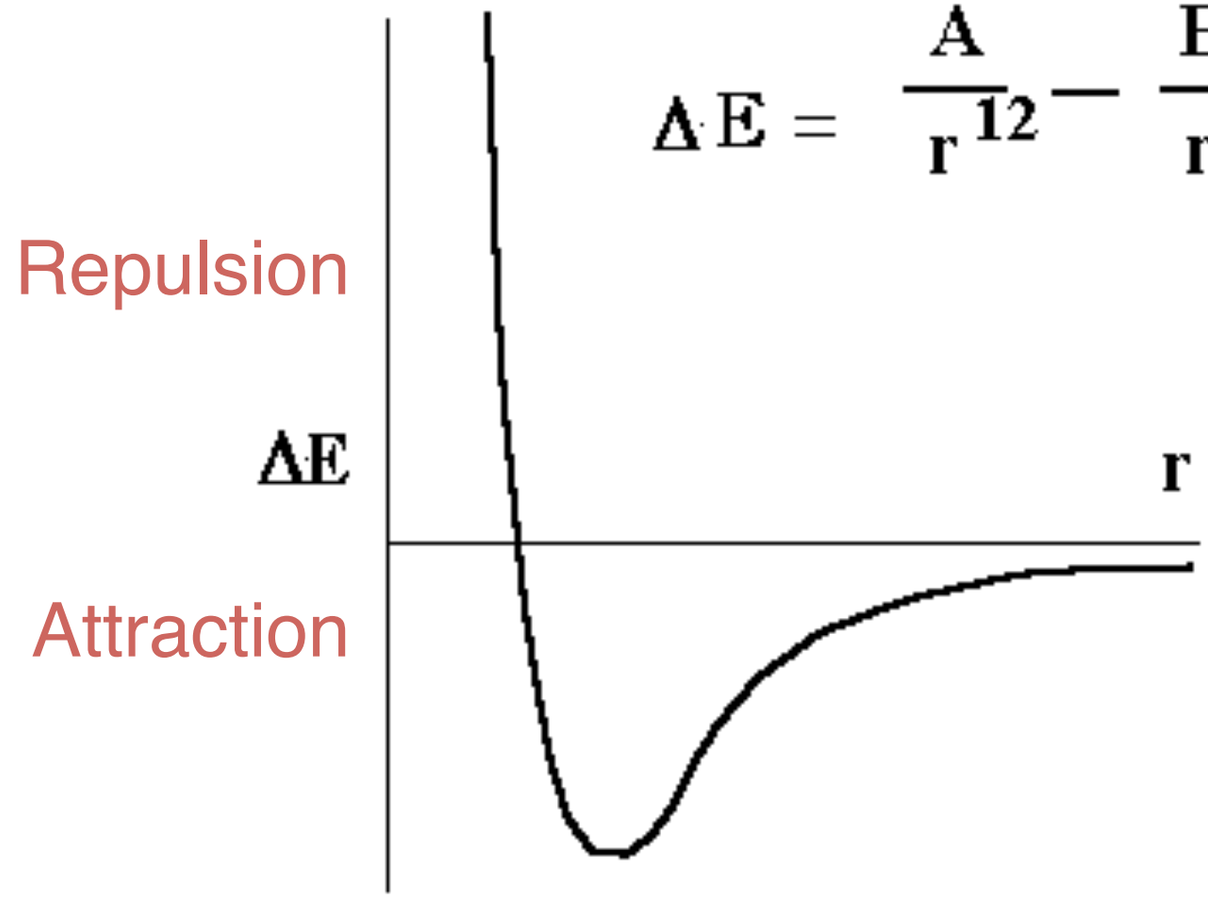
Key forces affecting structure:

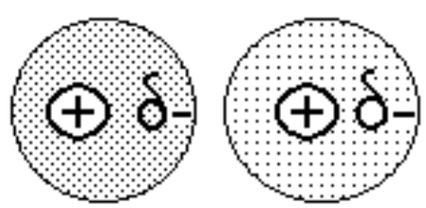
- H-bonding
- Van der Waals
- Electrostatics
- Hydrophobicity



Key forces affecting structure:

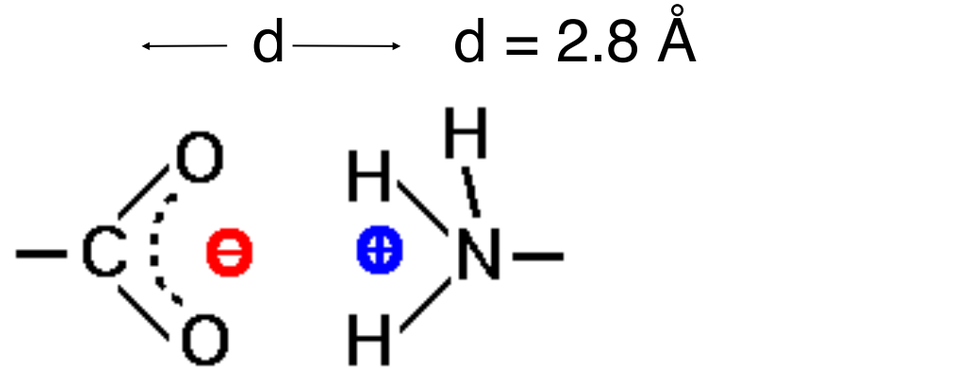
- H-bonding
- Van der Waals
- Electrostatics
- Hydrophobicity





Key forces affecting structure:

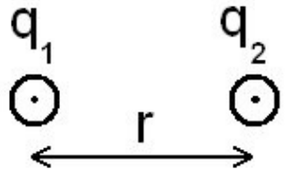
- H-bonding
- Van der Waals
- Electrostatics



carboxyl group and amino group

- Hydrophobicity

(some time called IONIC BONDs or SALT BRIDGEs)



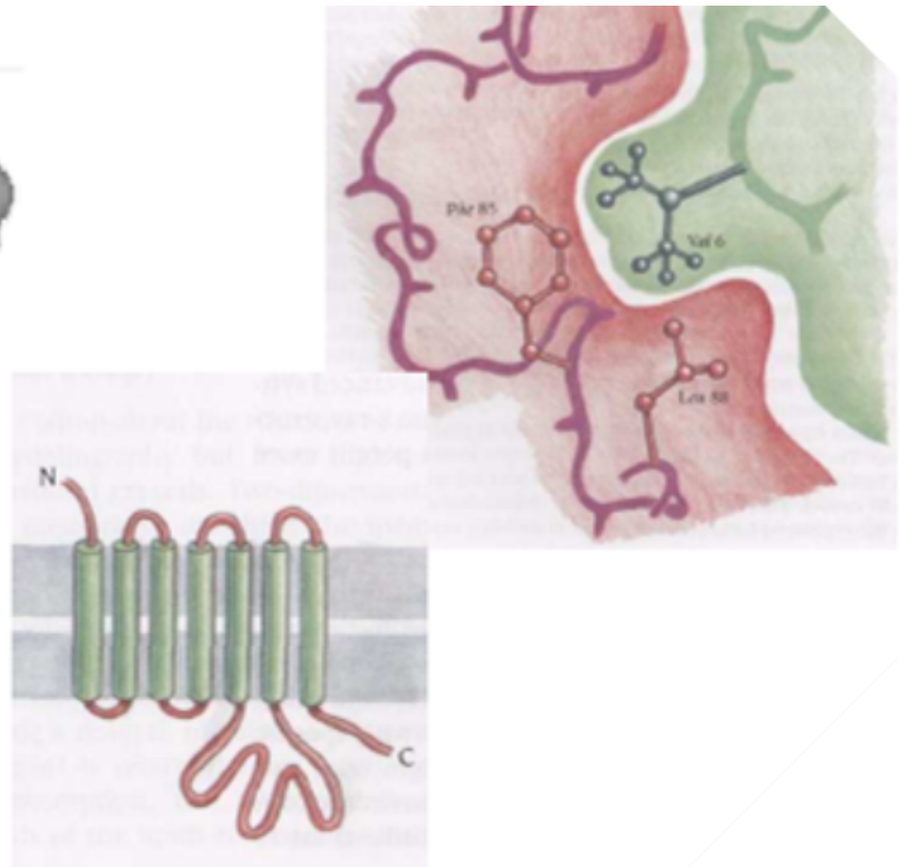
Coulomb's law

$$E = \frac{K\, q_1\, q_2}{D\, r}$$

E = Energy
k = constant
D = Dielectric constant (vacuum = 1; H_2O = 80)
q_1 & q_2 = electronic charges (Coulombs)
r = distance (Å)

Key forces affecting structure:

- H-bonding
- Van der Waals
- Electrostatics
- Hydrophobicity



The force that causes hydrophobic molecules or nonpolar portions of molecules to aggregate together rather than to dissolve in water is called Hydrophobicity (Greek, “water fearing”). This is not a separate bonding force; rather, it is the result of the energy required to insert a nonpolar molecule into water.

Today's Menu

- **Overview of structural bioinformatics**
  - Motivations, goals and challenges

- **Fundamentals of protein structure**
  - Structure composition, form and forces

- **Representing, interpreting & modeling protein structure**
  - Visualizing & interpreting protein structures
  - Analyzing protein structures
  - Modeling energy as a function of structure

Today's Menu

- **Overview of structural bioinformatics**
  - Motivations, goals and challenges

- **Fundamentals of protein structure**
  - Structure composition, form and forces

- **Representing, interpreting & modeling protein structure**
  - Visualizing & interpreting protein structures
  - Analyzing protein structures
  - Modeling energy as a function of structure

Do it Yourself!

Hand-on time!

Focus on **section 1** only please!

N.B. Remember to make your new **class11** RStudio project inside your GitHub tracked directory from last day and **UNCHECK** the "Create a Git repository" option...

Side-Note: PDB File Format

- PDB files contains atomic coordinates and associated information.

Amino Acid

Sequence/Residue Number

Element

Chain

Coordinates

X **Y** **Z**

(etc.)

```
ATOM      1  N   MET A   1      19.353  41.547  -3.887  ...
ATOM      2  CA  MET A   1      20.513  40.939  -4.592  ...
ATOM      3  C   MET A   1      20.150  39.658  -5.355  ...
ATOM      4  O   MET A   1      19.053  39.551  -5.903  ...
ATOM      5  CB  MET A   1      21.642  40.678  -3.592  ...
ATOM      6  CG  MET A   1      21.233  39.903  -2.360  ...
ATOM      7  SD  MET A   1      22.533  39.928  -1.113  ...
ATOM      8  CE  MET A   1      23.771  38.881  -1.885  ...
ATOM      9  N   ASP A   2      21.068  38.694  -5.390  ...
ATOM     10  CA  ASP A   2      20.856  37.440  -6.117  ...
ATOM     11  C   ASP A   2      20.124  36.371  -5.299  ...
ATOM     12  O   ASP A   2      20.680  35.818  -4.351  ...
```

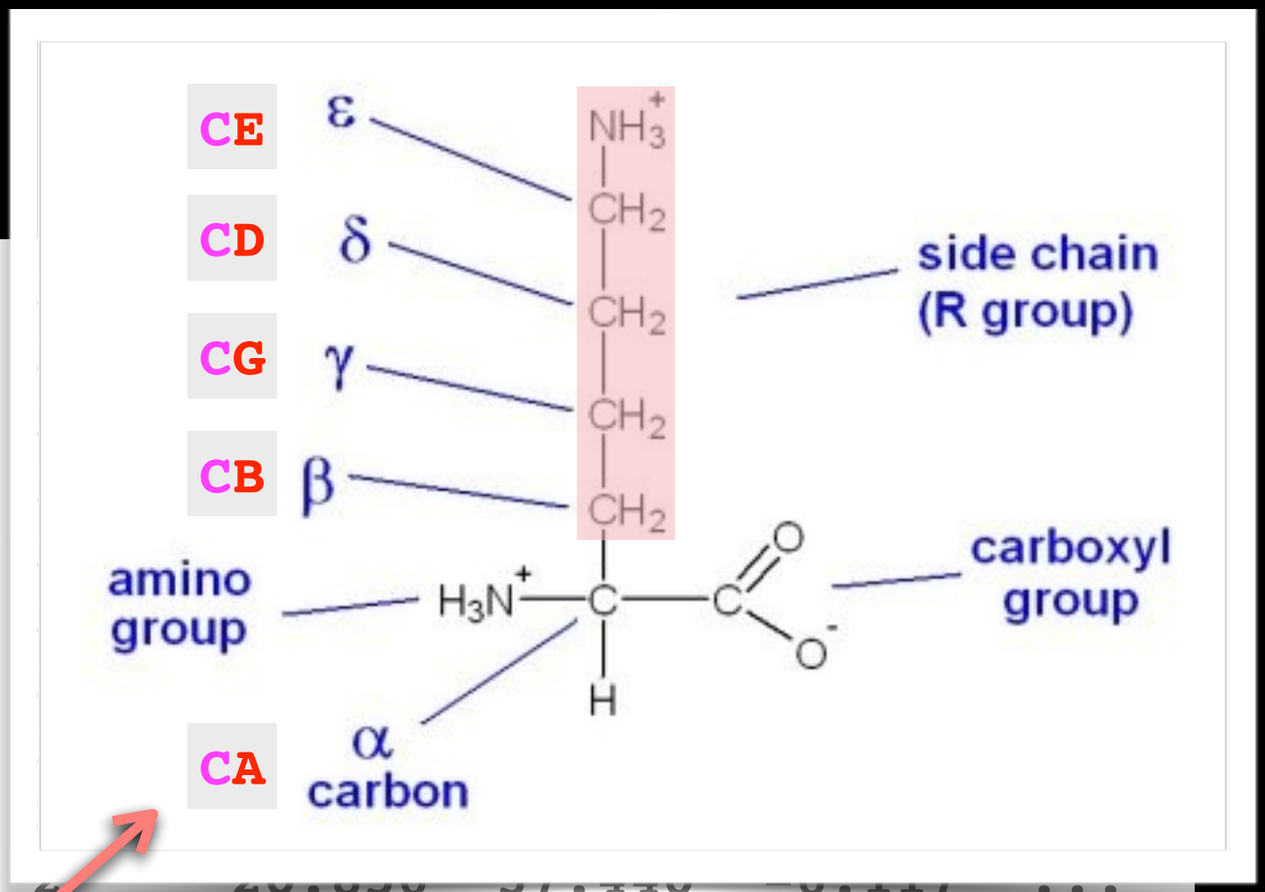
Element position within amino acid

Side-Note: PDB File Format

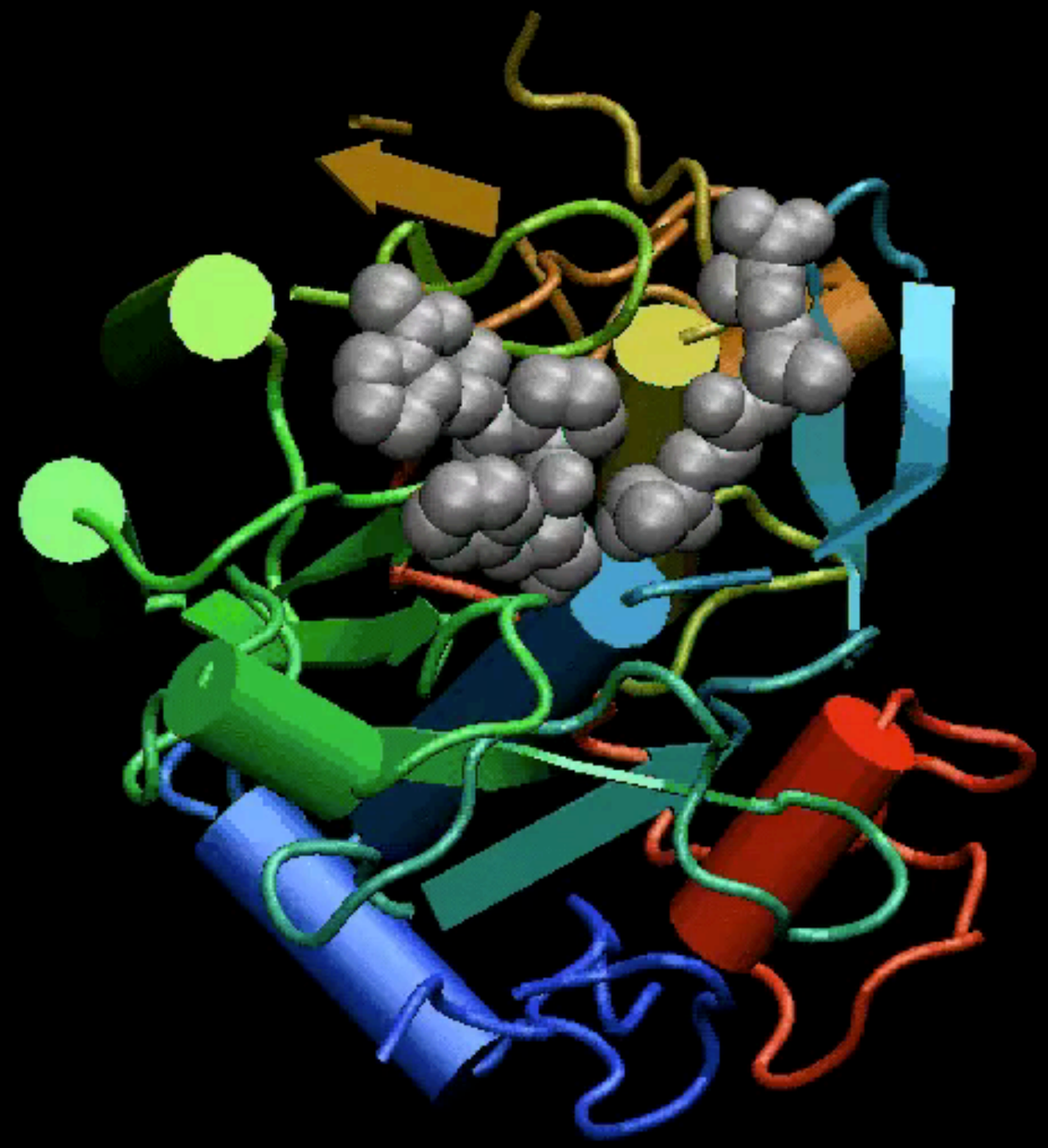
- PDB files contains atomic coordinates and associated information.

Element

```
ATOM      1  N    MET A
ATOM      2  CA   MET A
ATOM      3  C    MET A
ATOM      4  O    MET A
ATOM      5  CB   MET A
ATOM      6  CG   MET A
ATOM      7  SD   MET A
ATOM      8  CE   MET A
ATOM      9  N    ASP A
ATOM     10  CA   ASP A
ATOM     11  C    ASP A       2      20.124  36.371  -5.299  ...
ATOM     12  O    ASP A       2      20.680  35.818  -4.351  ...
```



Element position within amino acid



Download VMD

Hands-on Time!

Focus on **section 2** of "*Lab Sheet*" (using VMD)

Today's Menu

- **Overview of structural bioinformatics**
  - Motivations, goals and challenges

- **Fundamentals of protein structure**
  - Structure composition, form and forces

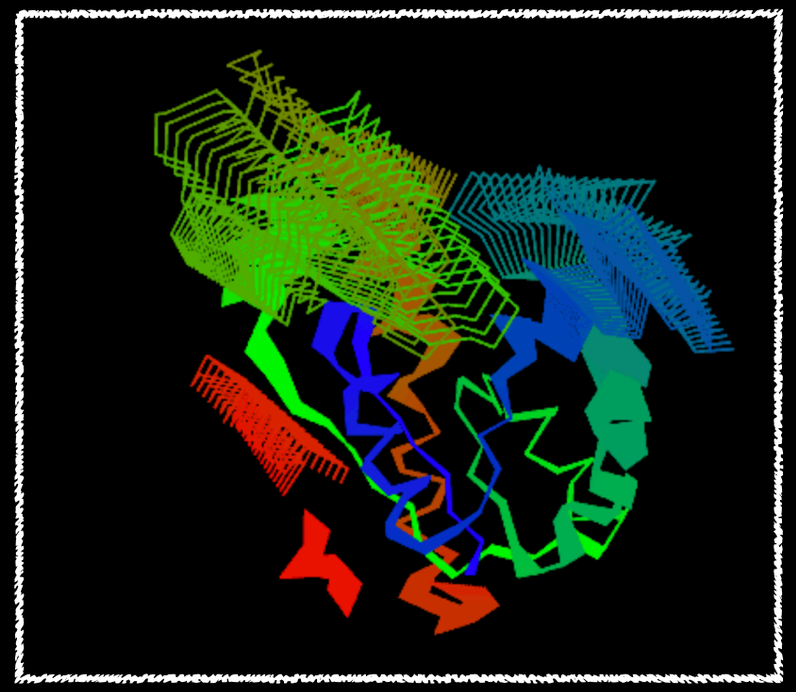
- **Representing, interpreting & modeling protein structure**
  - Visualizing and interpreting protein structures
  - Analyzing protein structures
  - Modeling energy as a function of structure

Do it Yourself!

Hand-on time!

Focus on **section 3** to **5**

Bio3D view()



- If you want the 3D viewer in your R markdown you can install the development version of bio3d.view

- In your R console:

```
> install.packages("devtools")
> devtools::install_bitbucket("Grantlab/bio3d-view")
```

- To use in your R session:

```
> library("bio3d.view")
> pdb <- read.pdb("5p21")
> view(pdb)
> view(pdb, "overview", col="sse")
```

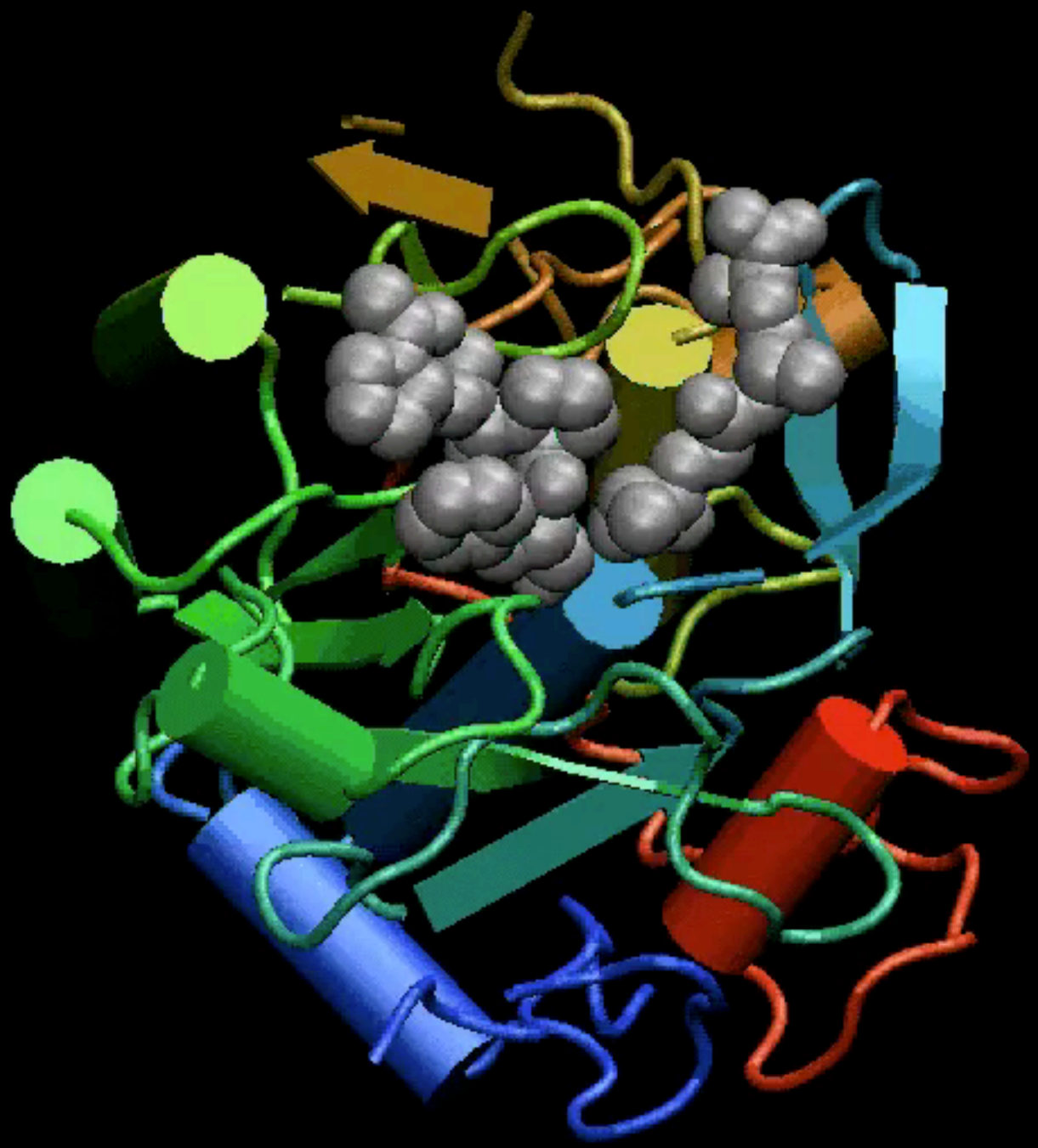
Today's Menu

- **Overview of structural bioinformatics**
  - Motivations, goals and challenges

- **Fundamentals of protein structure**
  - Structure composition, form and forces

- **Representing, interpreting & modeling protein structure**
  - Visualizing and interpreting protein structures
  - Analyzing protein structures
  - Modeling energy as a function of structure

NMA models the protein as a network of elastic strings



Proteinase K

NMA in Bio3D

- Normal Mode Analysis (NMA) is a bioinformatics method that can predict the major motions of biomolecules.

```{r}
library(bio3d)
library(bio3d.view)
```

```{r}
pdb <- read.pdb("1hel")
modes <- nma( pdb )
m7 <- mktrj(modes, mode=7, file="mode_7.pdb")

view(m7, col=vec2color(rmsf(m7)))
```

Bio3D view()

- If you want the interactive 3D viewer in **Rmd** rendered to output: html_output document:

```
```{r}
library(bio3d.view)
library(rgl)
```
```

```
```{r}
modes <- nma( read.pdb("1hel") )
m7 <- mktrj(modes, mode=7, file="mode_7.pdb")

view(m7, col=vec2color(rmsf(m7)))
rglwidget(width=500, height=500)
```
```

KEY CONCEPT: POTENTIAL FUNCTIONS DESCRIBE A SYSTEMS **ENERGY** AS A FUNCTION OF ITS **STRUCTURE**

Two main approaches:

(1). Physics-Based

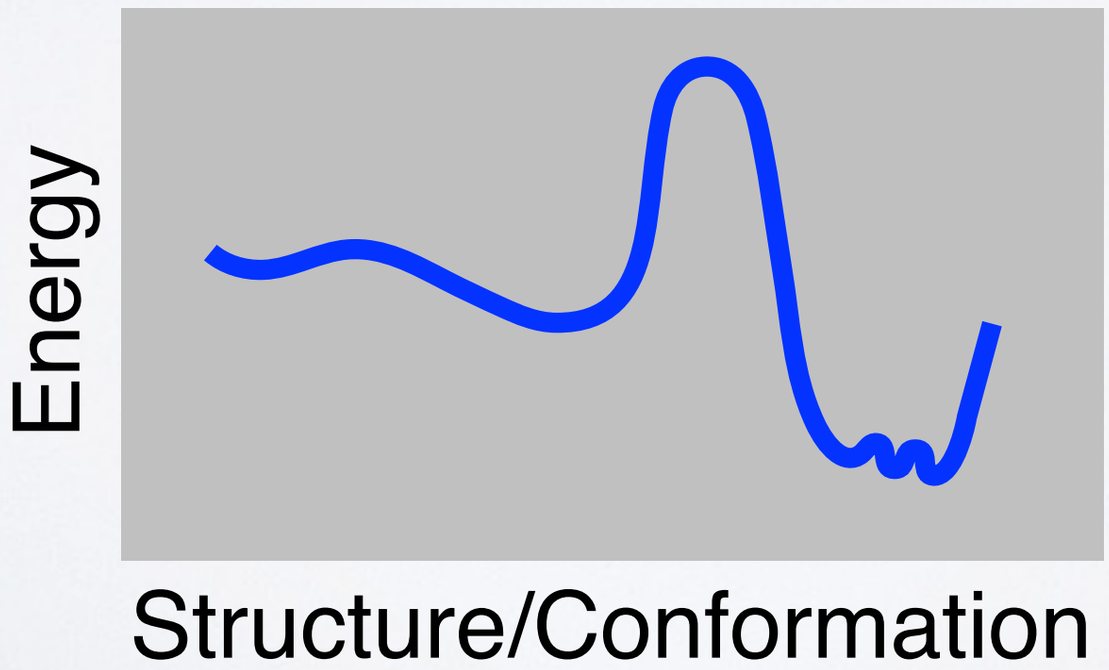
(2). Knowledge-Based

KEY CONCEPT: POTENTIAL FUNCTIONS DESCRIBE A SYSTEMS **ENERGY** AS A FUNCTION OF ITS **STRUCTURE**

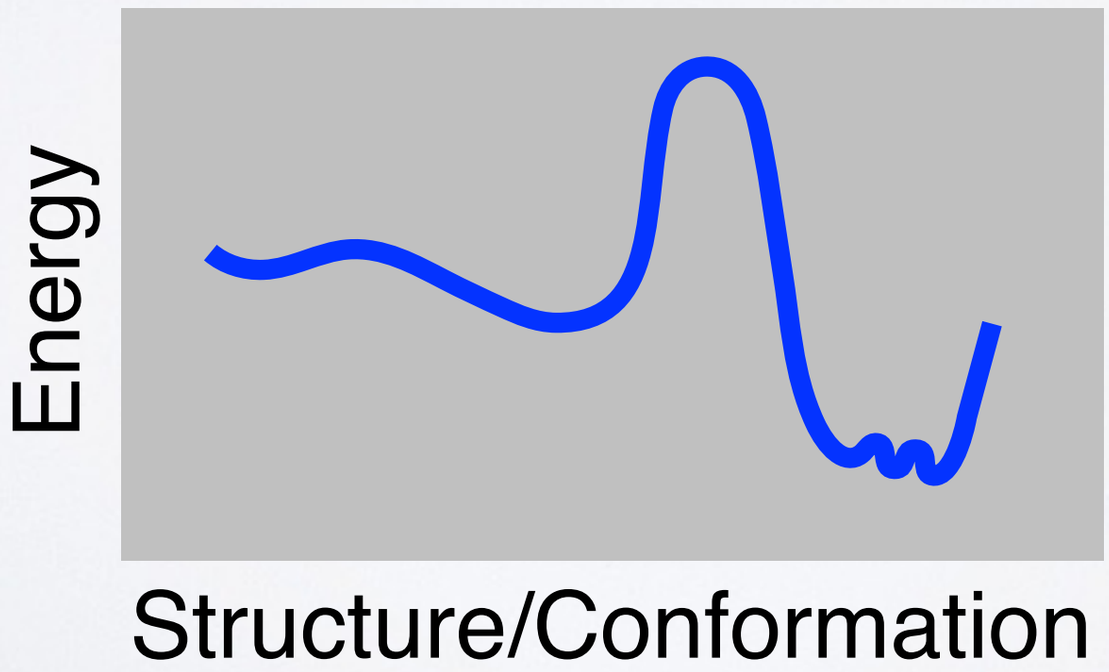
Two main approaches:

(1). Physics-Based

(2). Knowledge-Based



This will be the focus of the next class!



SUMMARY

- Structural bioinformatics is computer aided structural biology
- Described major motivations, goals and challenges of structural bioinformatics
- Reviewed the fundamentals of protein structure
- Explored how to use R to perform advanced custom structural bioinformatics analysis!
- Introduced both physics and knowledge based modeling approaches for describing the structure, energetics and dynamics of proteins computationally

[Muddy Point Assessment]