

# **STRUCTURAL BIOINFORMATICS**

Barry Grant  
**University of Michigan**

[www.thegrantlab.org](http://www.thegrantlab.org)

[bjgrant@umich.edu](mailto:bjgrant@umich.edu)

## Objective:

Provide an introduction to the practice of structural bioinformatics, major goals, current research challenges, and application areas.

## **Q. What does Bioinformatics mean to you?**

*“Bioinformatics is the application of computers to the collection, archiving, organization, and interpretation of biological data.” [Orengo, 2003]*

- ... Bioinformatics is a hybrid of biology and computer science
- ... **Bioinformatics is computer aided biology!**

## **Q. So what is STRUCTURAL bioinformatics?**

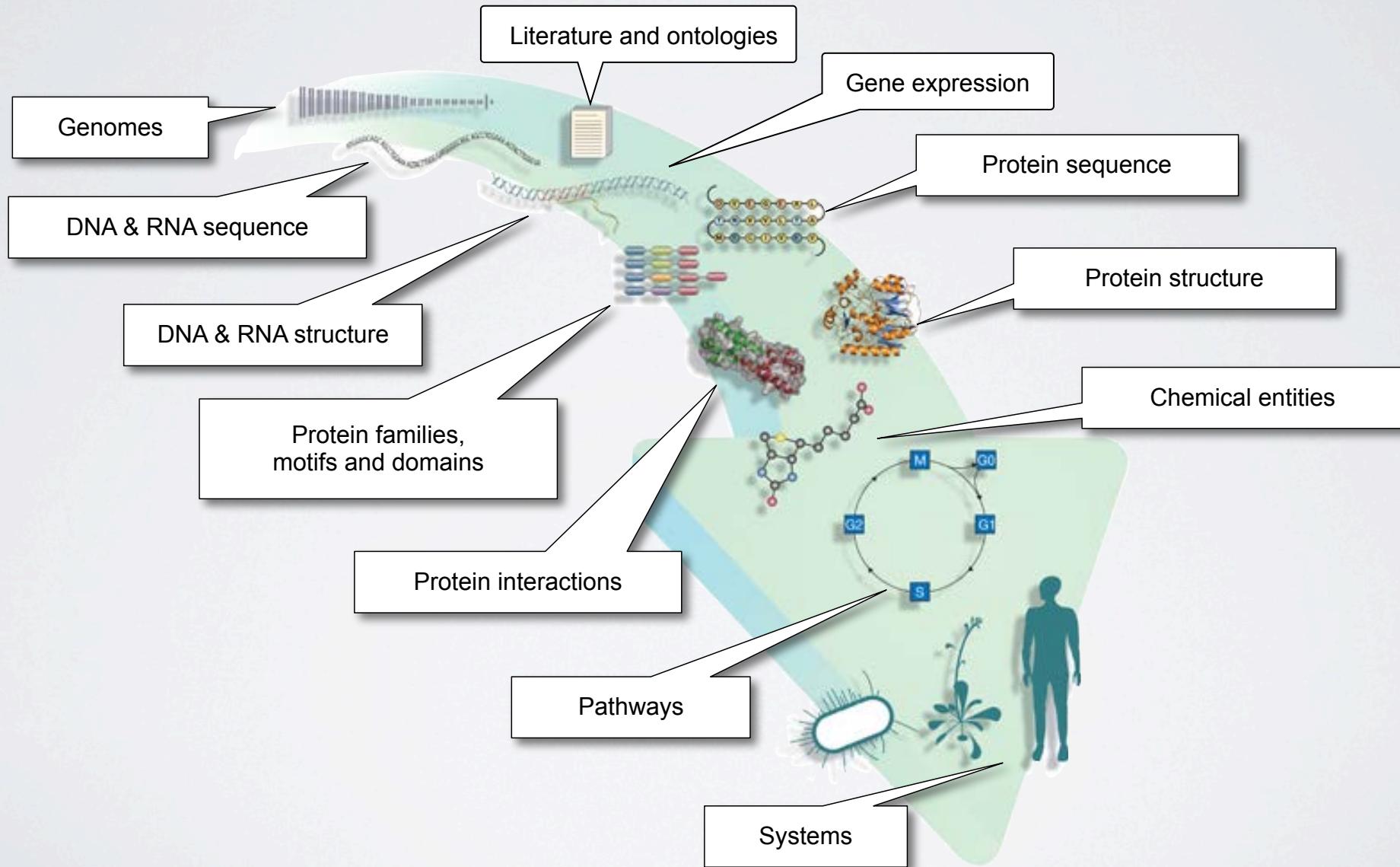
- **Structural bioinformatics is computer aided structural biology!**
- Characterizes biomolecules and their assemblies at the molecular & atomic level.

## **Q. Why should we care?**

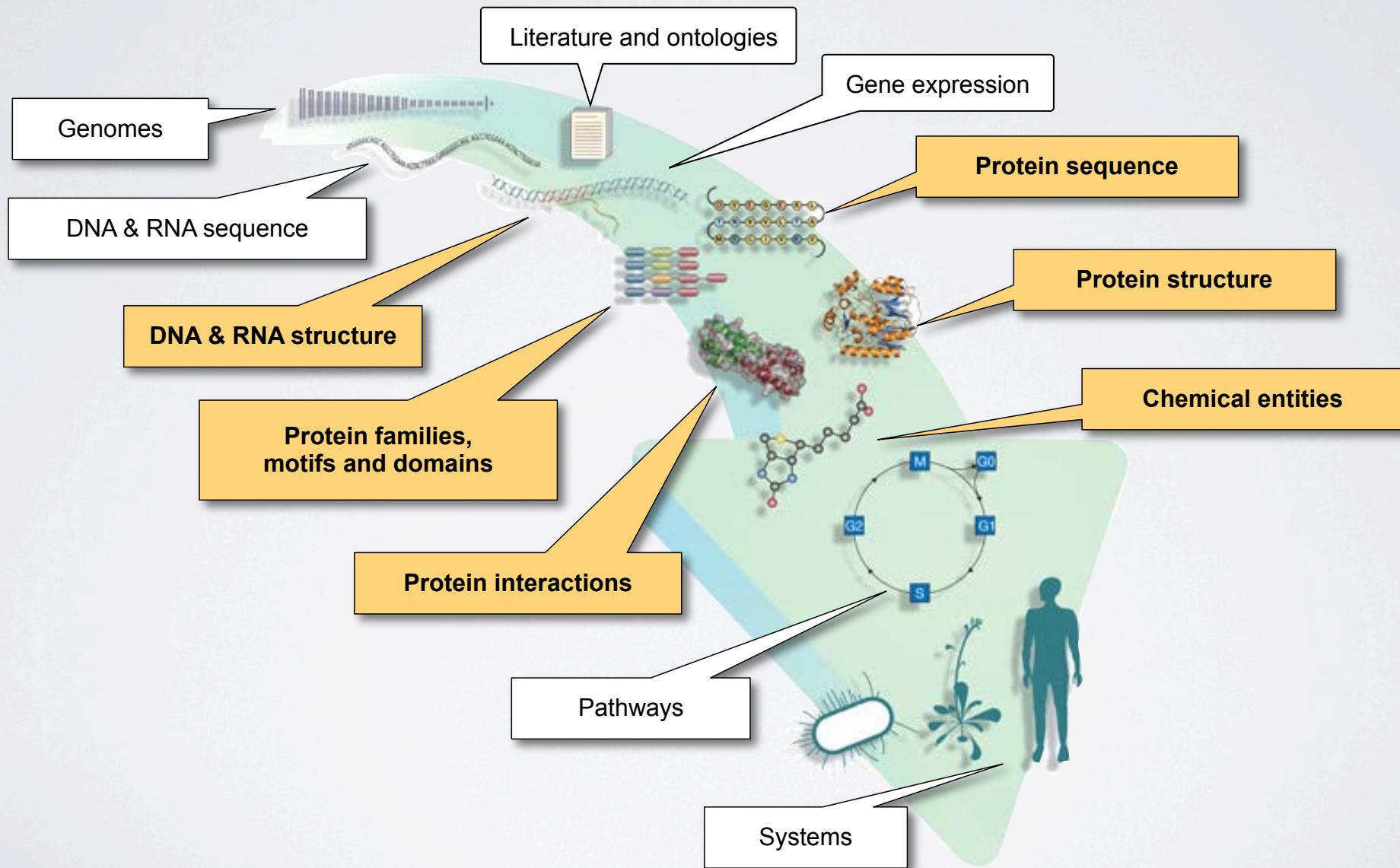
- Because biomolecules are “nature’s robots” [Tanford, 2001]

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

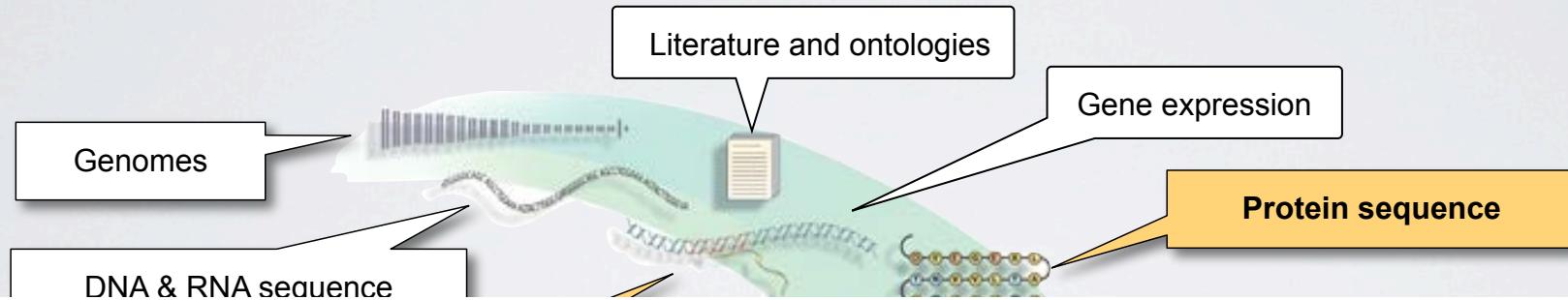
# BIOINFORMATICS DATA



# STRUCTURAL DATA IS CENTRAL

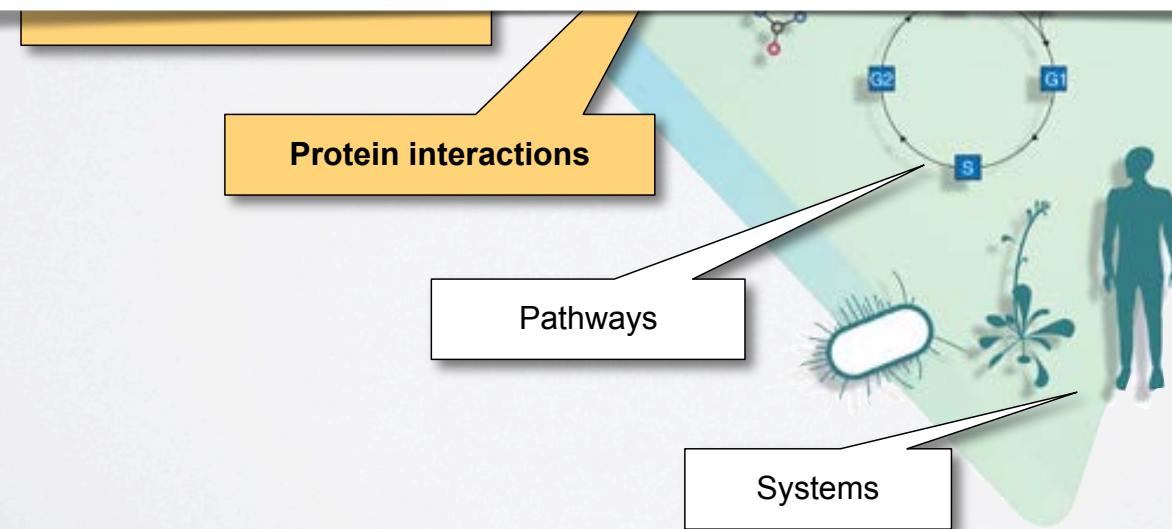


# STRUCTURAL DATA IS CENTRAL



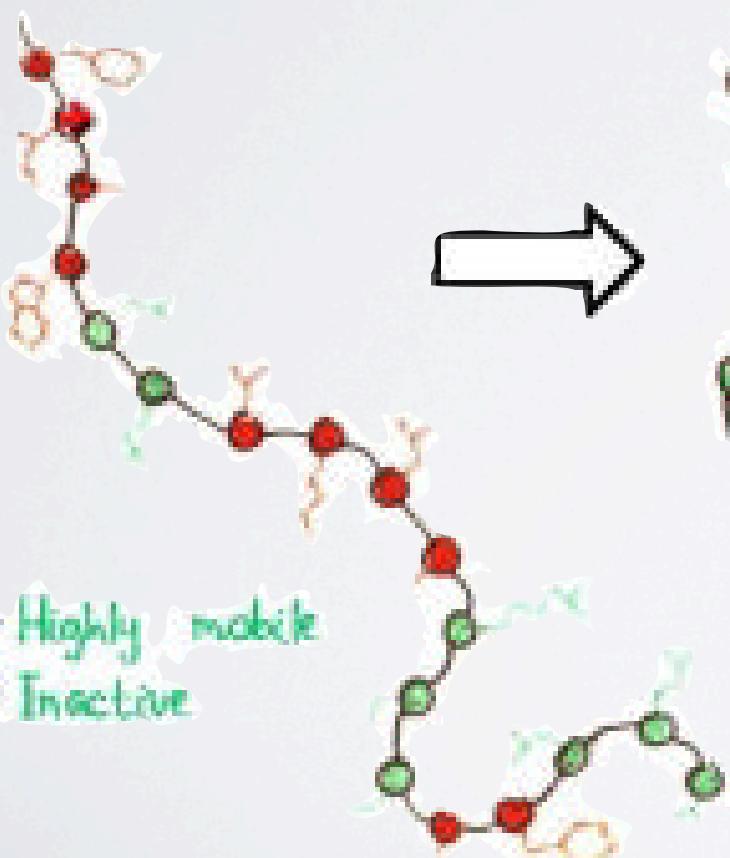
## THE HOLY TRINITY OF STRUCTURAL BIOINFORMATICS

**Sequence > Structure > Function**



# Sequence > Structure > Function

- Unfolded protein is a chain of amino acids



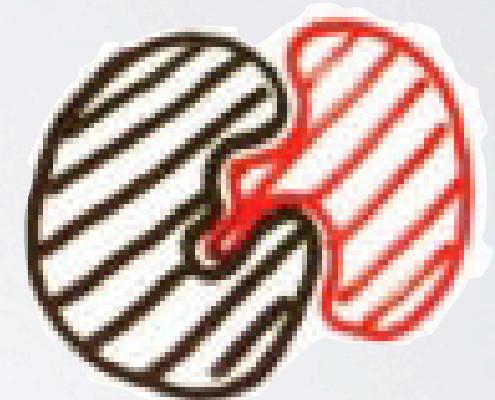
- Highly mobile
- Inactive

- Folded protein



- Unique shape
- Precisely ordered
- Stable
- Active

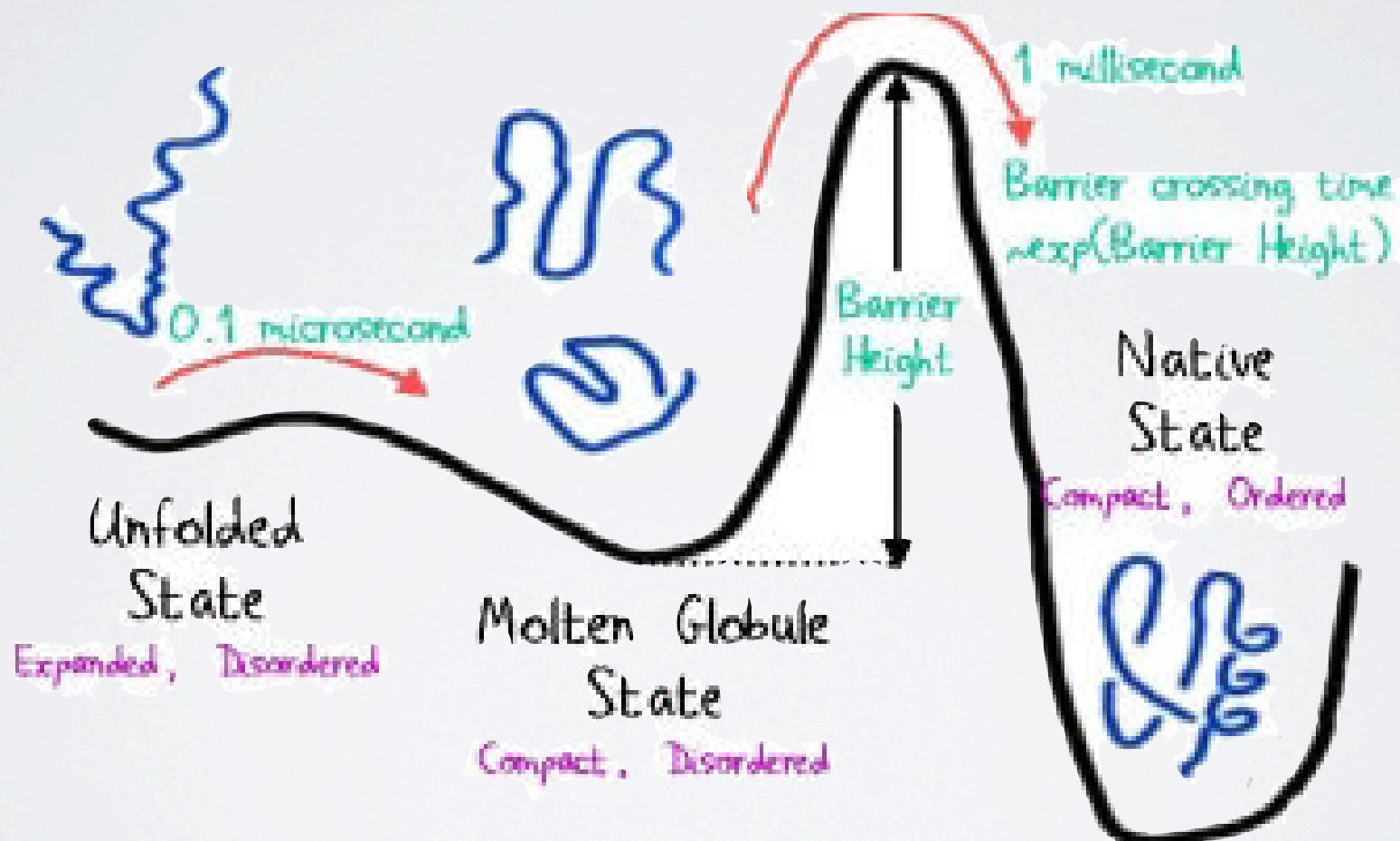
- Function depends on protein shape



- Specific associations
- Precise reactions

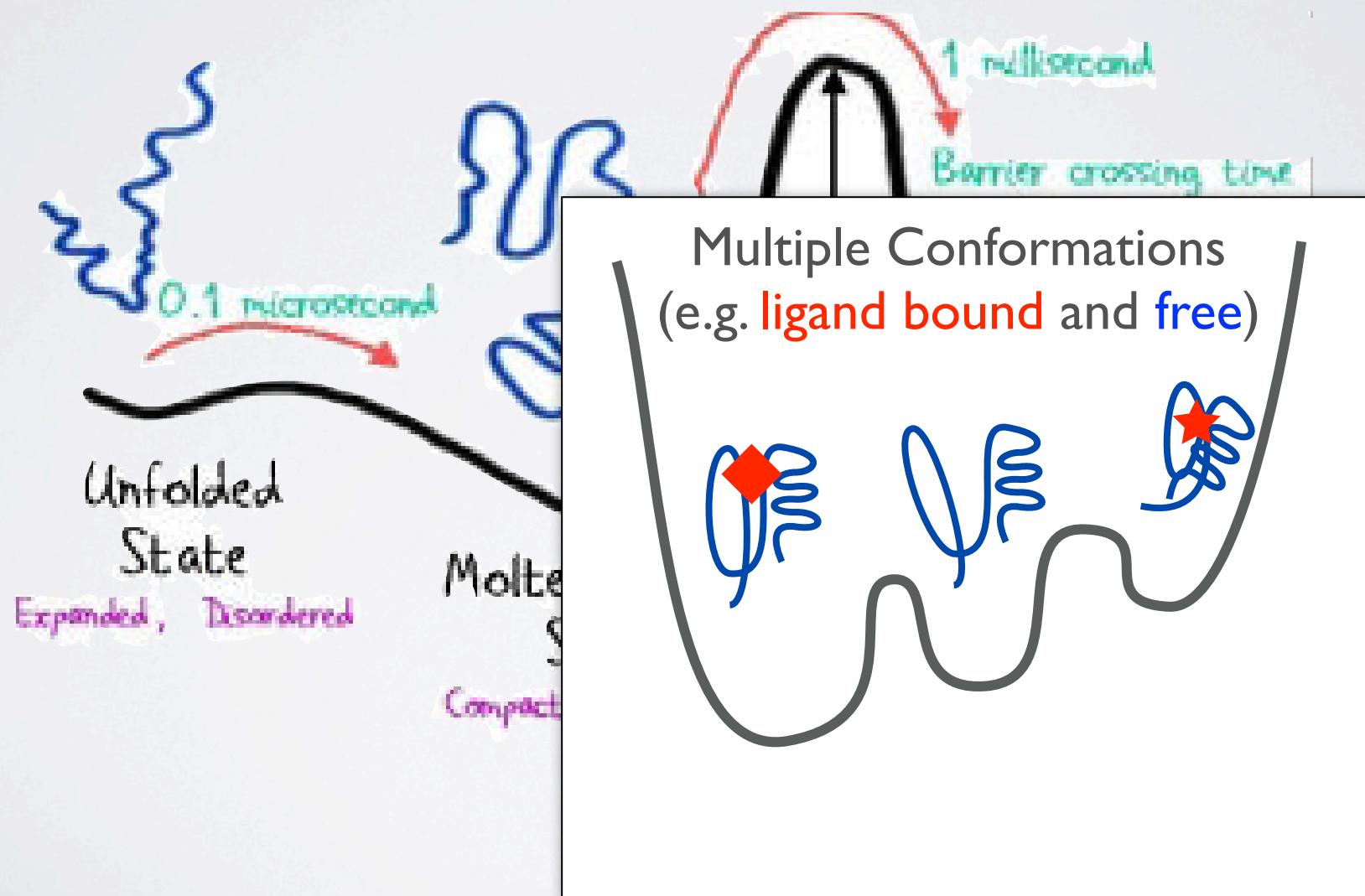
Slide Credit: Michael Levitt

# KEY CONCEPT: ENERGY LANDSCAPE



Slide Credit: Michael Levitt

# KEY CONCEPT: ENERGY LANDSCAPE



# TODAY'S MENU:

- **Overview of structural bioinformatics**
  - Motivations, Goals and Challenges
- **Fundamentals of protein structure**
  - Structure composition, form and forces
- **Representing and interpreting biomolecular structure**
  - PDB and SCOP databases
  - Modeling energy as a function of structure
    - Physics based and knowledge based approaches
- **Example Application Areas**
  - Structure based drug discovery
    - Receptor and ligand based approaches
  - Predicting functional dynamics
    - Molecular dynamics and normal mode analysis
  - Protein structure and function prediction

# TODAY'S MENU:

- **Overview of structural bioinformatics**
  - Motivations, Goals and Challenges
- **Fundamentals of protein structure**
  - Structure composition, form and forces
- **Representations**

## Next Lecture:

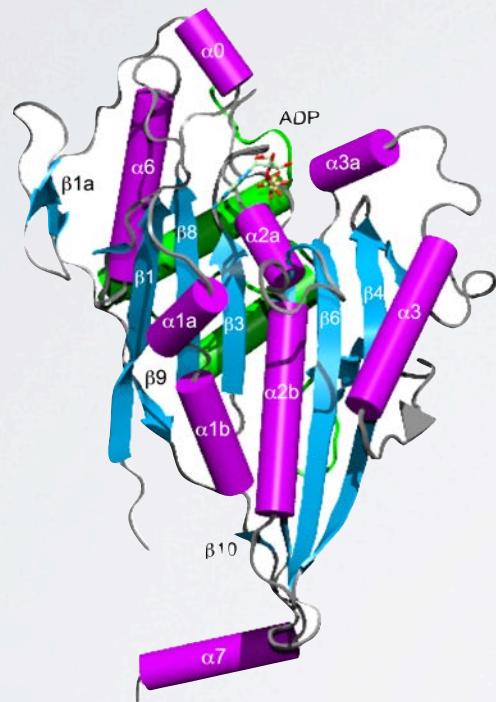
- Predicting structure from sequence [Prof. Zhang]
  - Structure prediction
  - Sequence based and knowledge based approaches

- **Example Application Areas**
  - Structure based drug discovery
    - Receptor and ligand based approaches
  - Predicting functional dynamics
    - Molecular dynamics and normal mode analysis
  - Protein structure and function prediction

# TODAY'S MENU:

- **Overview of structural bioinformatics**
  - Motivations, Goals and Challenges
- **Fundamentals of protein structure**
  - Structure composition, form and forces
- **Representing and interpreting biomolecular structure**
  - PDB and SCOP databases
  - Modeling energy as a function of structure
    - Physics based and knowledge based approaches
- **Example Application Areas**
  - Structure based drug discovery
    - Receptor and ligand based approaches
  - Predicting functional dynamics
    - Molecular dynamics and normal mode analysis
  - Protein structure and function prediction

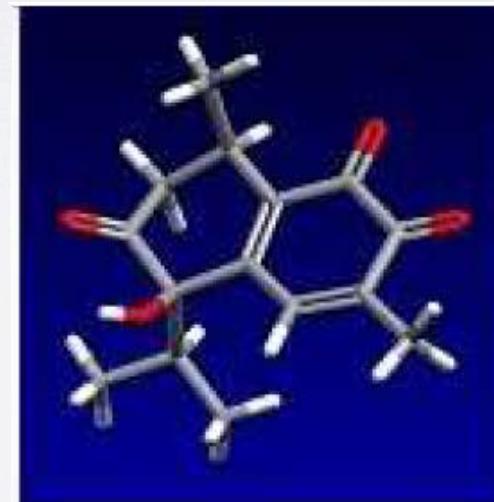
# TRADITIONAL FOCUS **PROTEIN**, **DNA** AND **SMALL MOLECULE** DATA SETS WITH **MOLECULAR STRUCTURE**



Protein  
(PDB)



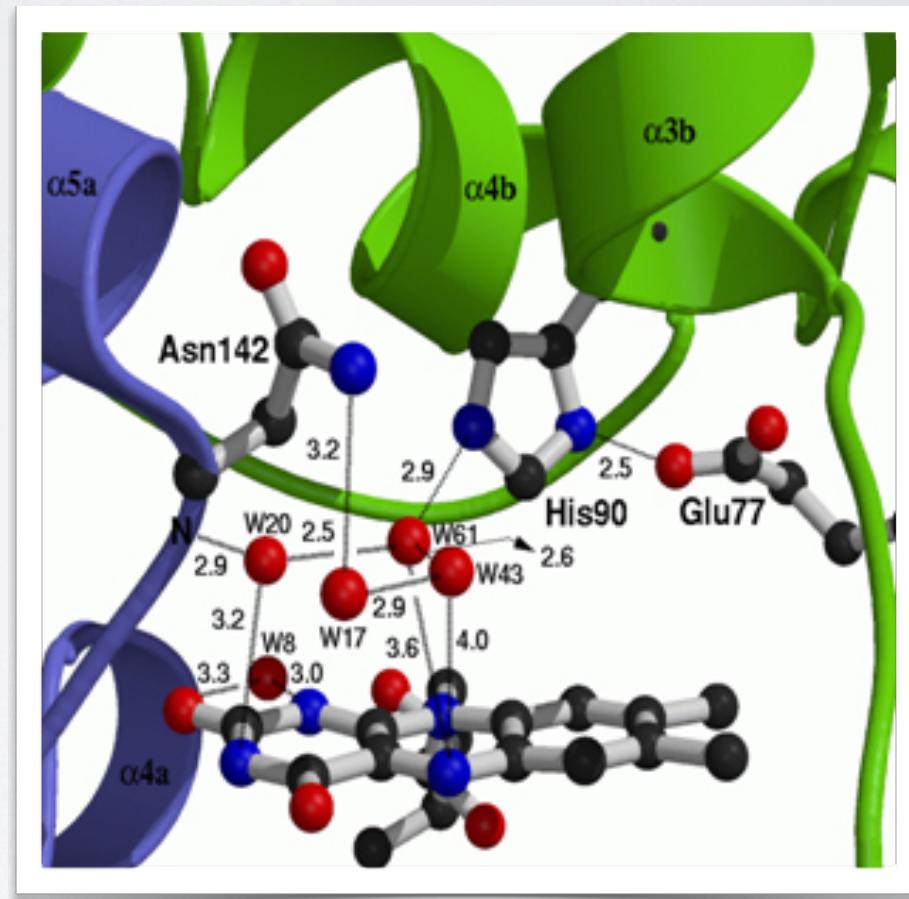
DNA  
(NDB)



Small Molecules  
(CCDB)

## Motivation 1: Detailed understanding of molecular interactions

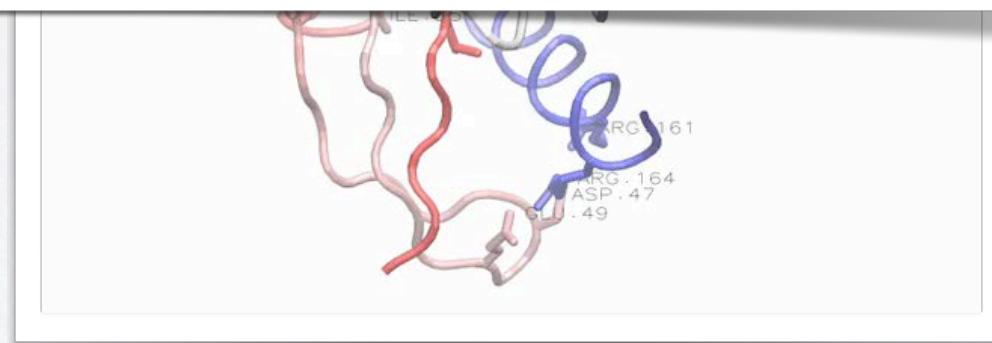
Provides an invaluable structural context for conservation and mechanistic analysis leading to functional insight.



## Motivation 1:

**Energetics      Dynamics**  
**Sequence ^ Structure ^ Function**

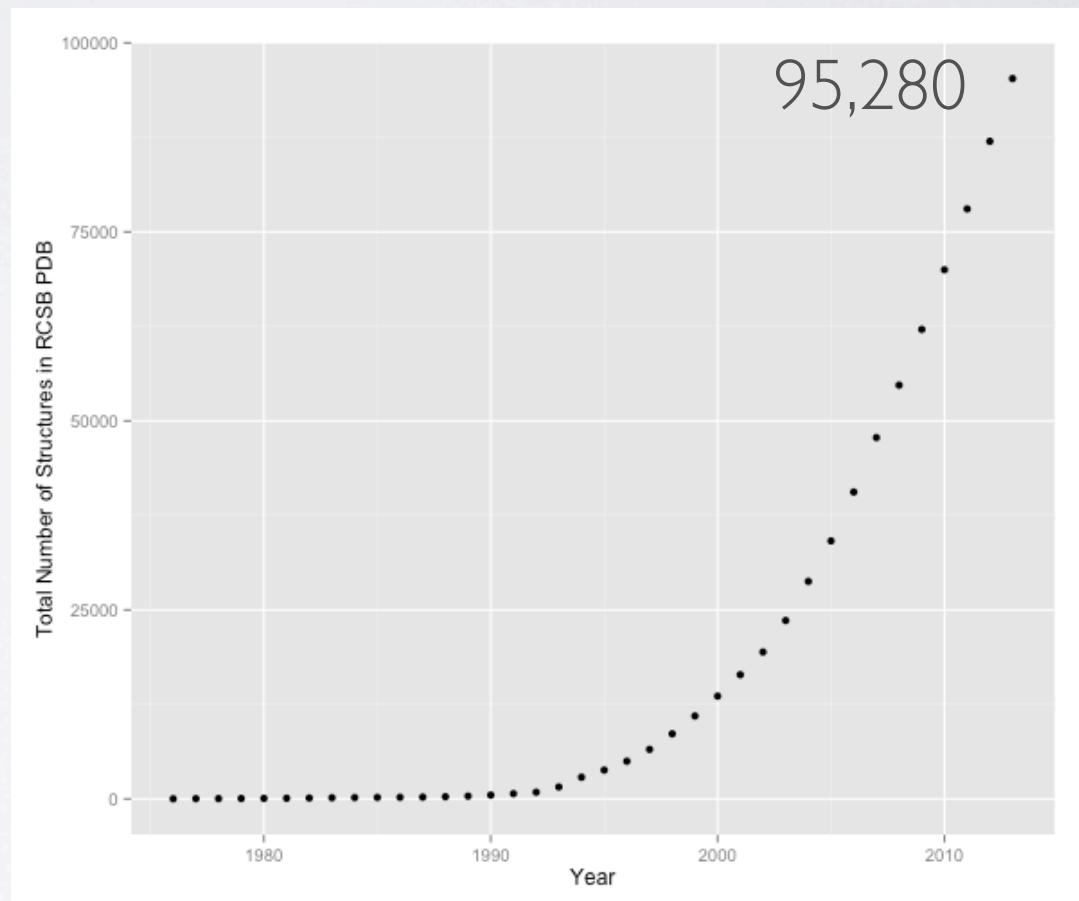
Computational modeling can provide detailed insight into functional interactions, their regulation and potential consequences of perturbation.



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: <http://www.rcsb.org/pdb/statistics/>

## Motivation 2: Lots of structural data is becoming available

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

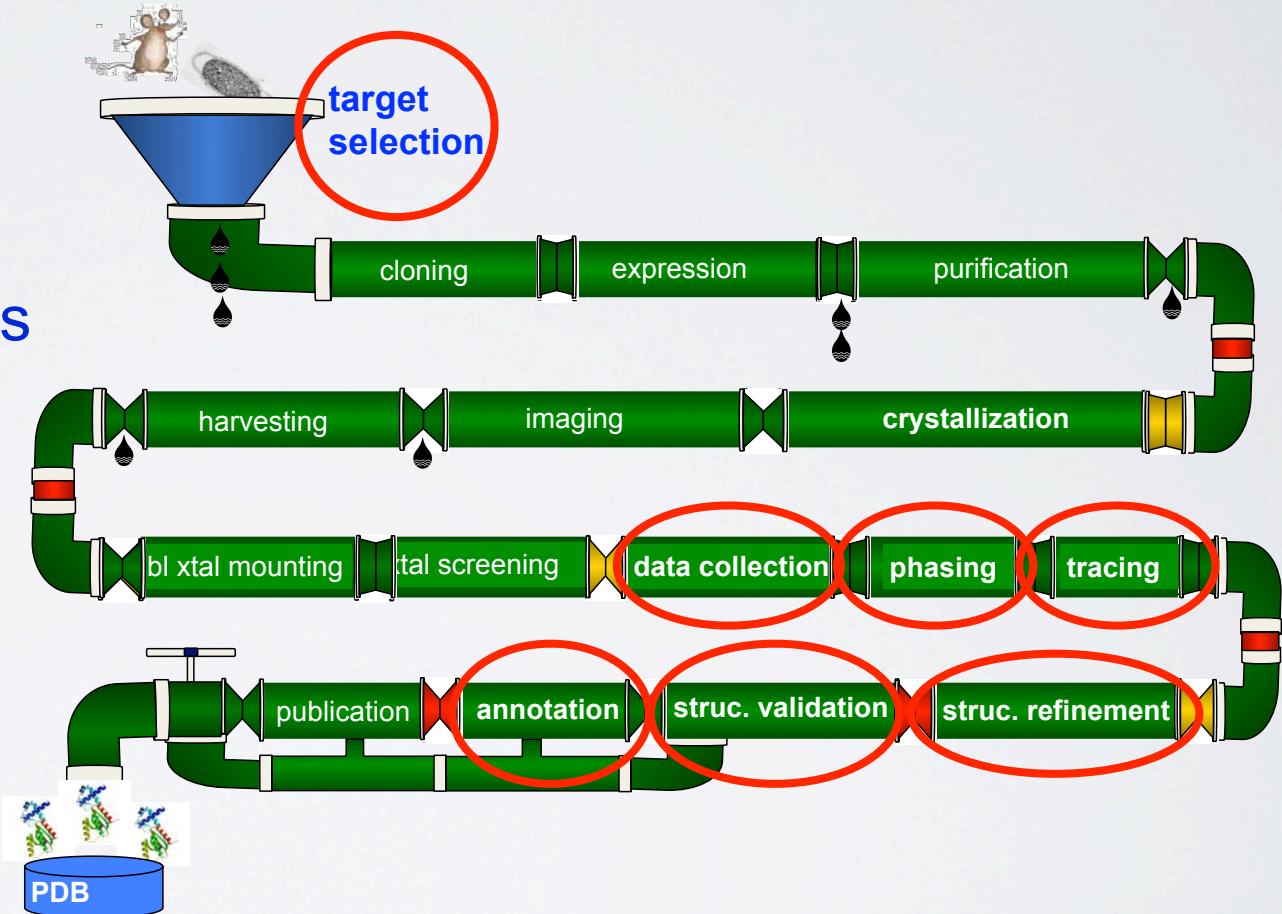
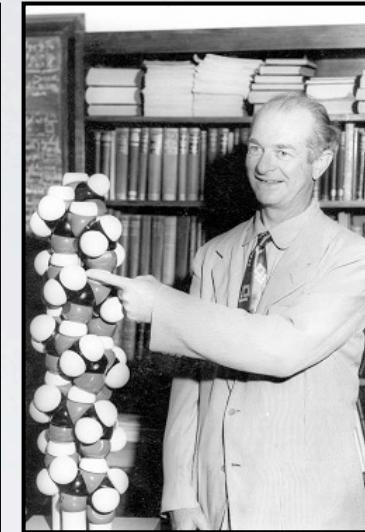
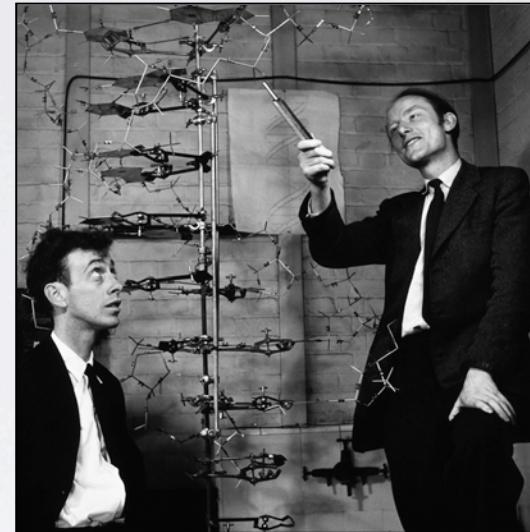


Image Credit: "Structure determination assembly line" Adam Godzik

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**

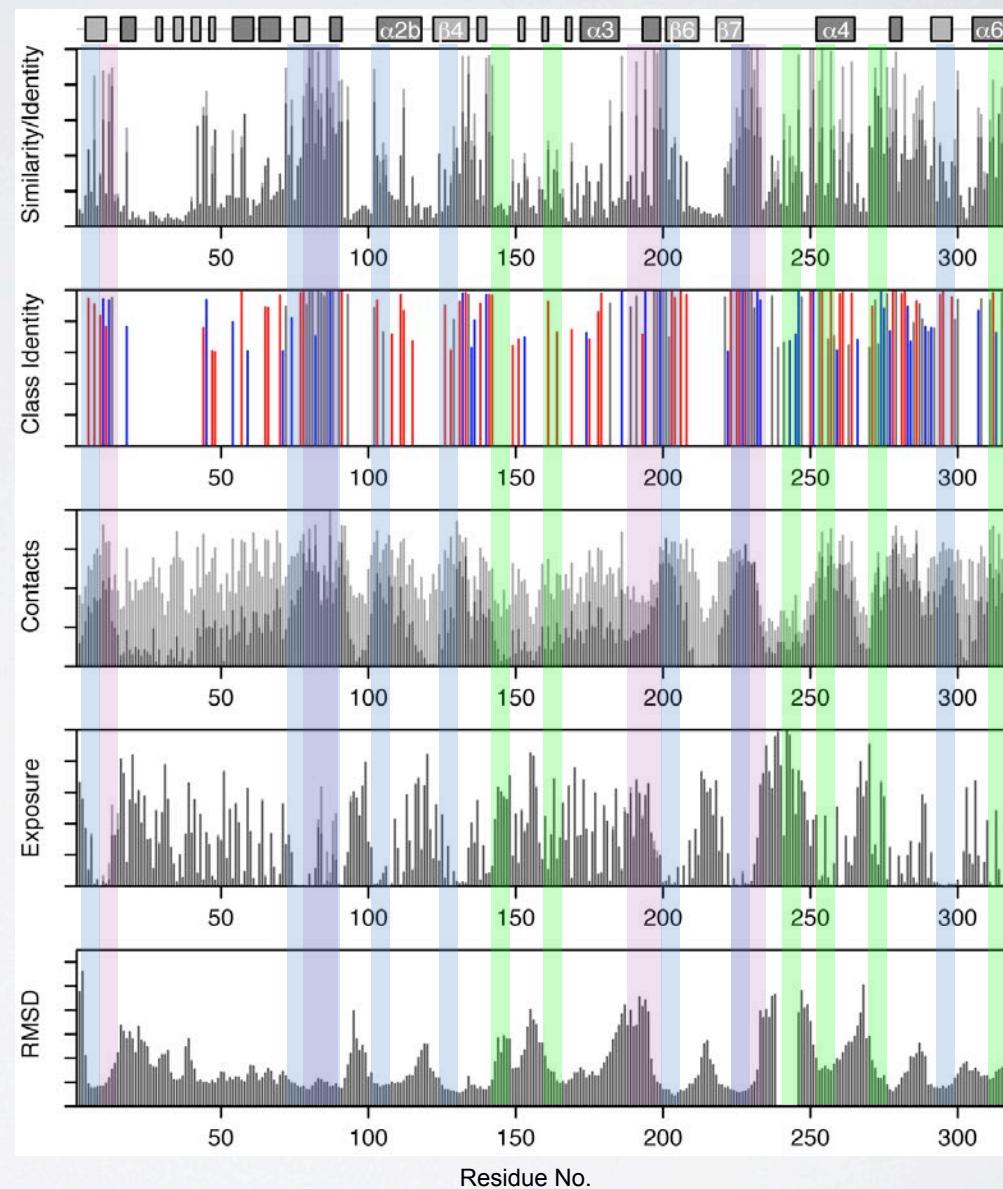
- Structure allows identification of more distant evolutionary relationships

## **Structure is encoded in sequence**

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

## Goals:

- Analysis
- Visualization
- Comparison
- Prediction
- Design



Grant et al. JMB. (2007)

## Goals:

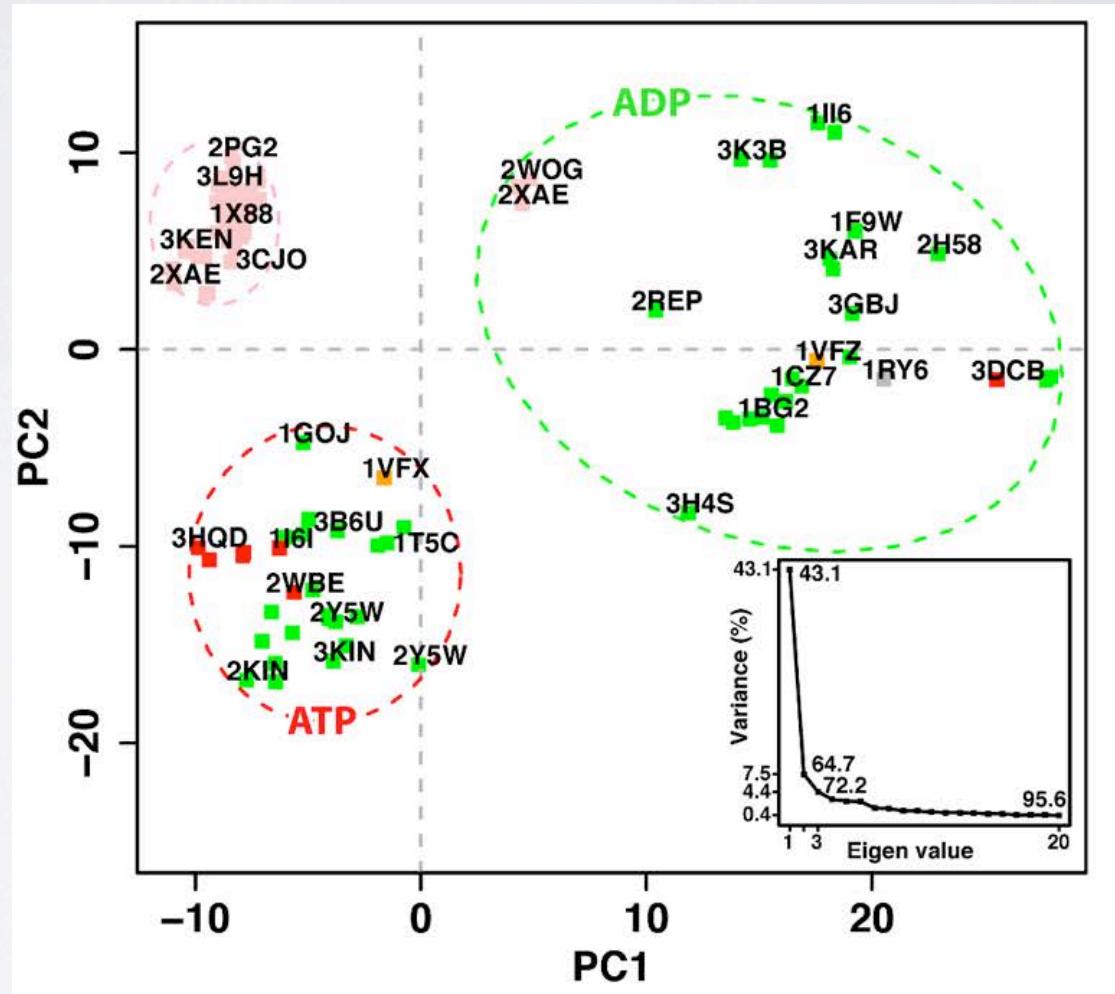
- Analysis
- Visualization
- Comparison
- Prediction
- Design



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

## Goals:

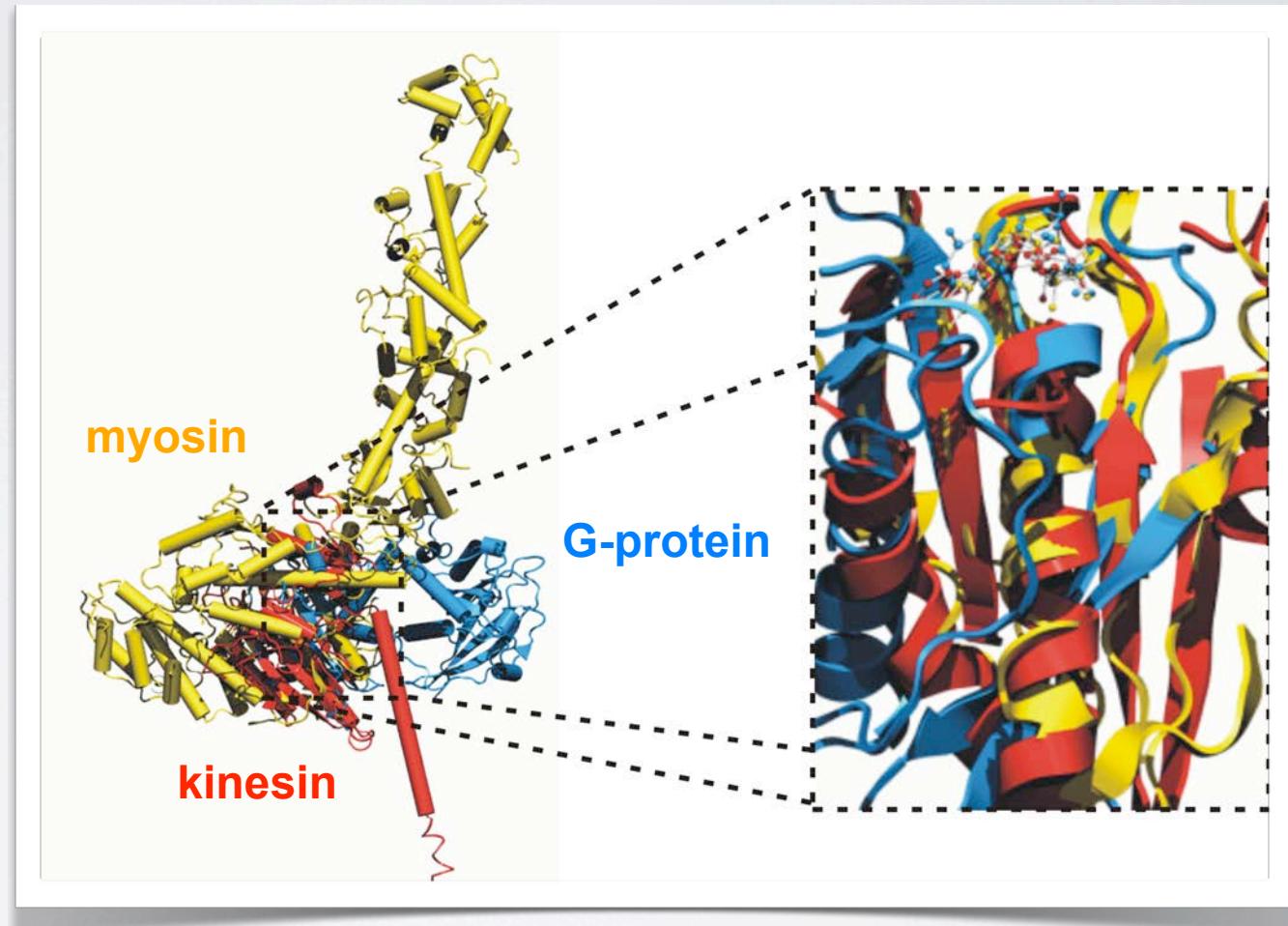
- Analysis
- Visualization
- Comparison
- Prediction
- Design



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

## Goals:

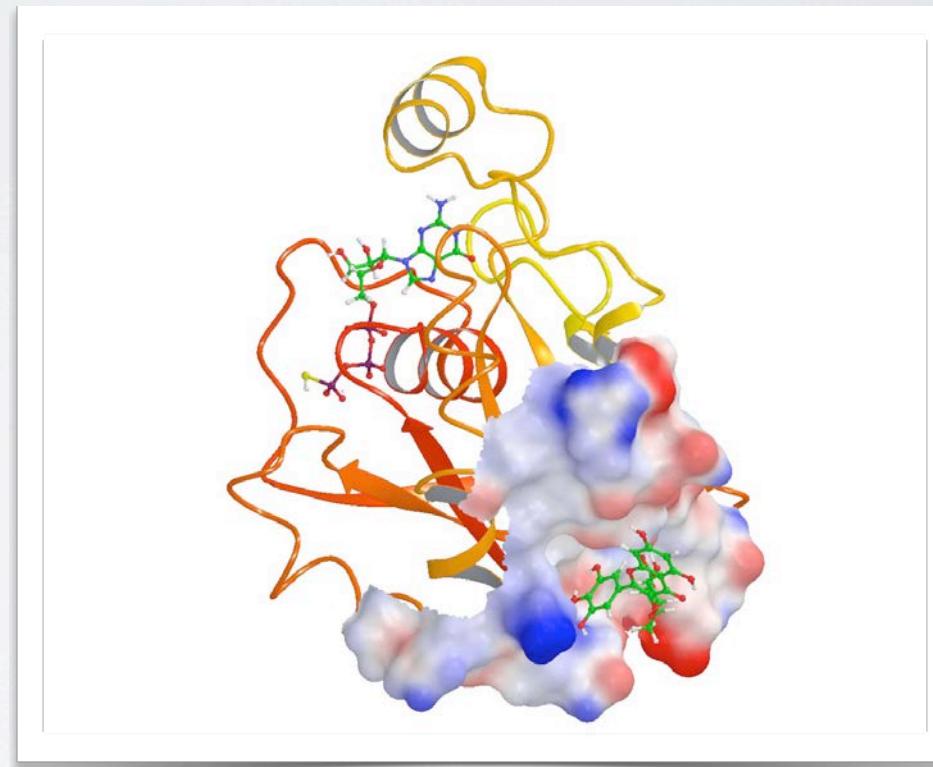
- Analysis
- Visualization
- Comparison
- Prediction
- Design



Grant et al. unpublished

## Goals:

- Analysis
- Visualization
- Comparison
- Prediction
- Design



Grant et al. PLoS One (2011, 2012)

## Goals:

- Analysis
- Visualization
- 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

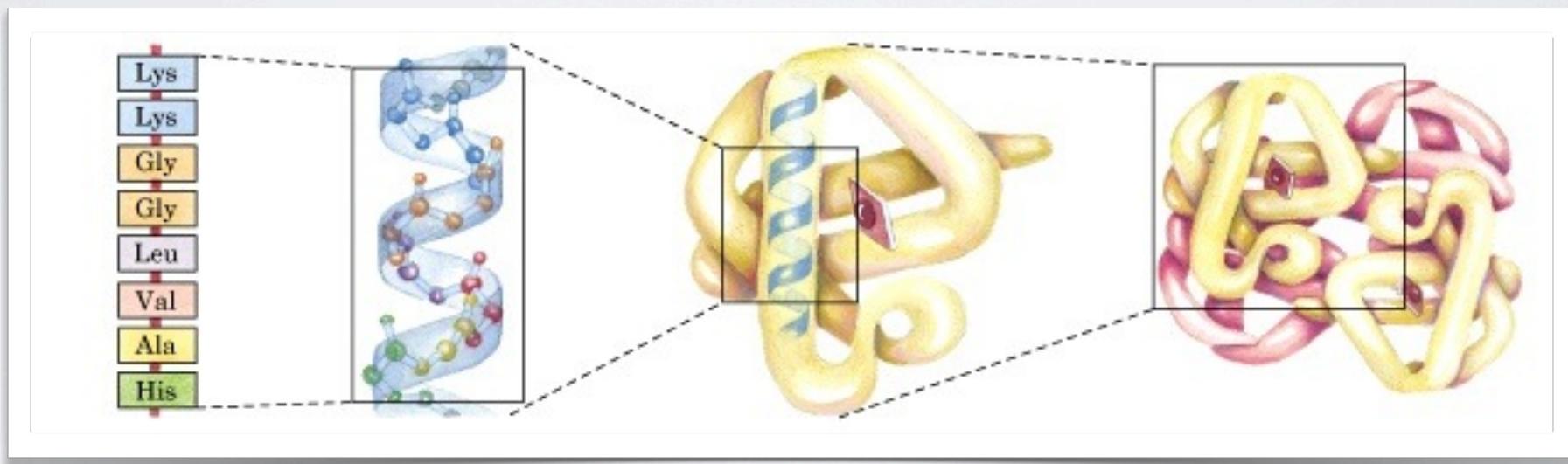
...BREAK...

# TODAY'S MENU:

- **Overview of structural bioinformatics**
  - Motivations, Goals and Challenges
- **Fundamentals of protein structure**
  - Structure composition, form and forces
- **Representing and interpreting biomolecular structure**
  - PDB and SCOP databases
  - Modeling energy as a function of structure
    - Physics based and knowledge based approaches
- **Example Application Areas**
  - Structure based drug discovery
    - Receptor and ligand based approaches
  - Predicting functional dynamics
    - Molecular dynamics and normal mode analysis
  - Protein structure and function prediction

# 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

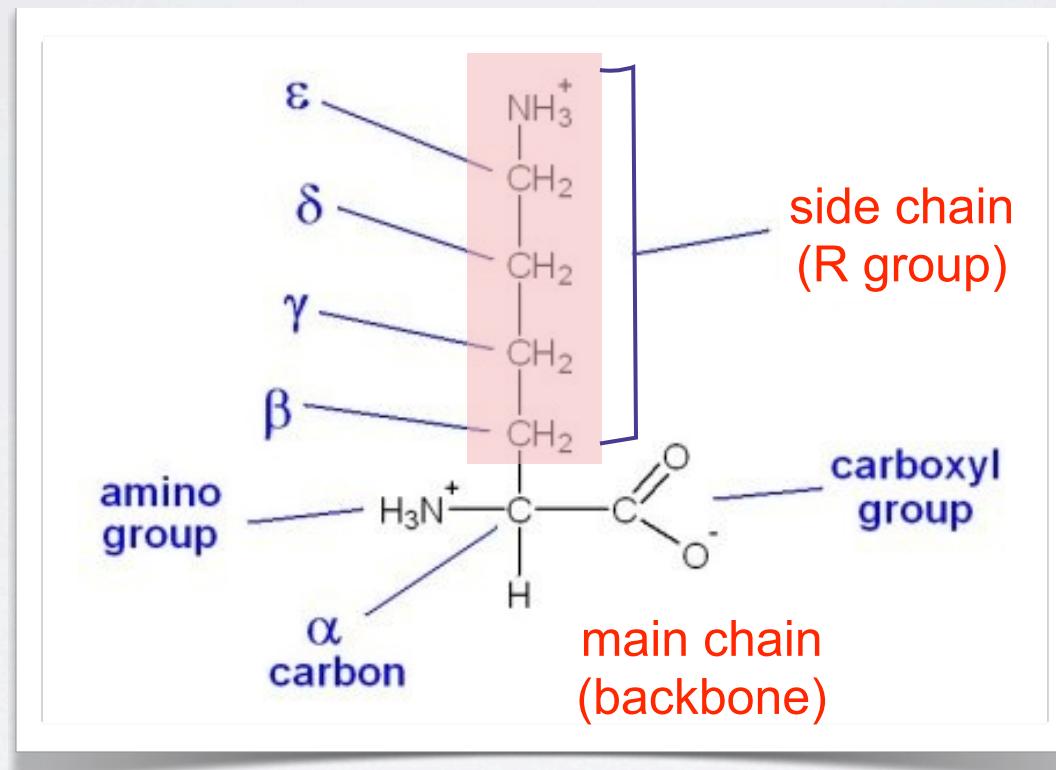
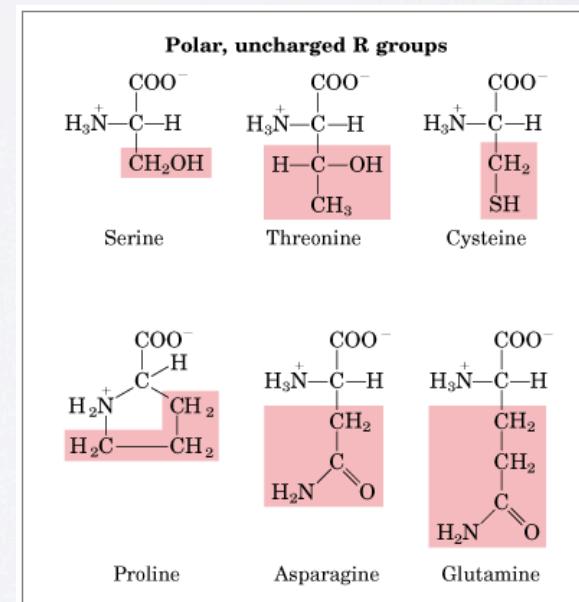
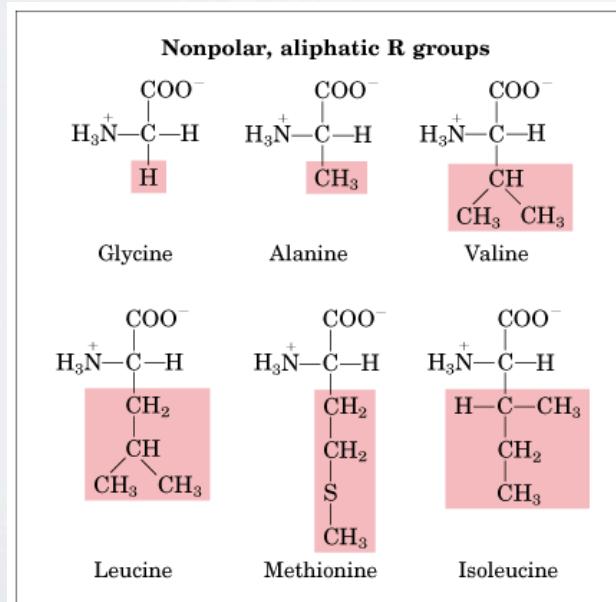
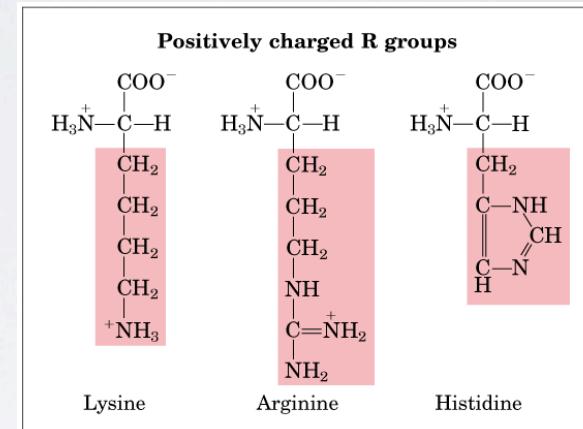
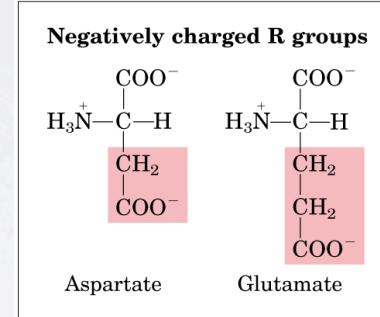
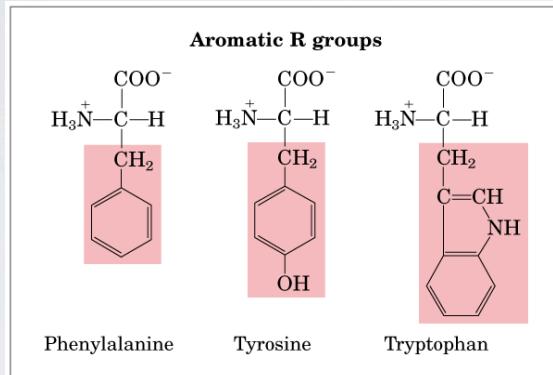


Image from: <http://www.ncbi.nlm.nih.gov/books/NBK21581/>

# AMINO ACIDS CAN BE GROUPED BY THE PHYSIOCHEMICAL PROPERTIES



# AMINO ACIDS POLYMERIZE THROUGH PEPTIDE BOND FORMATION

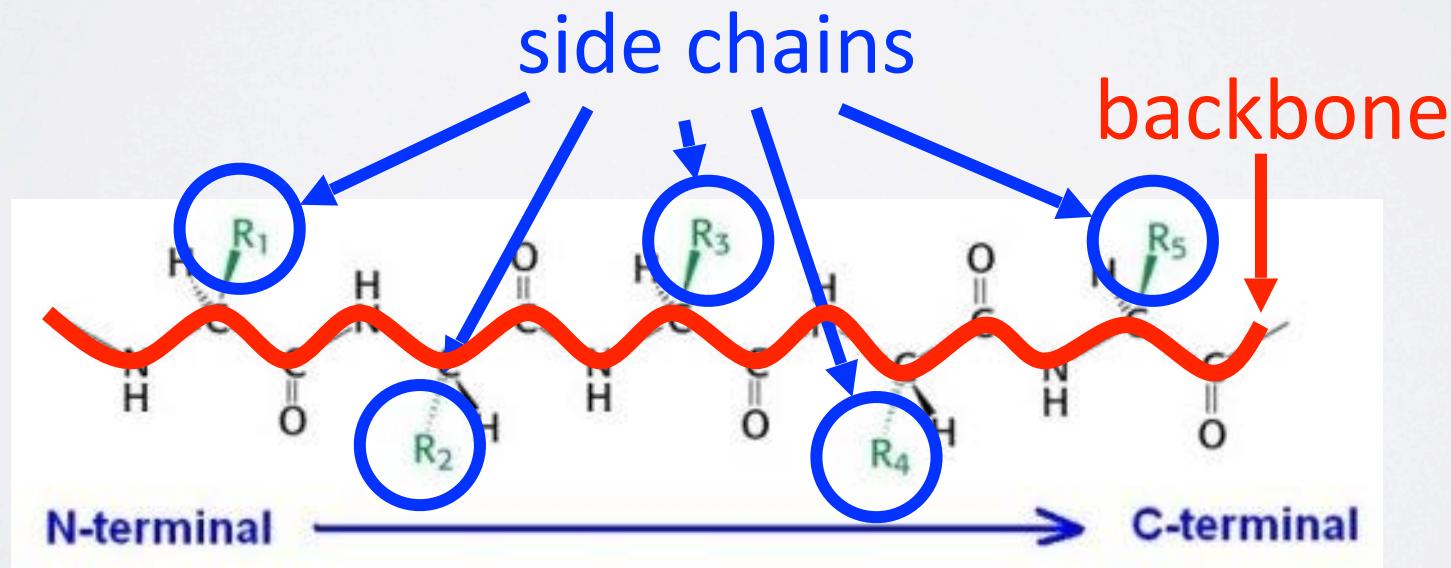
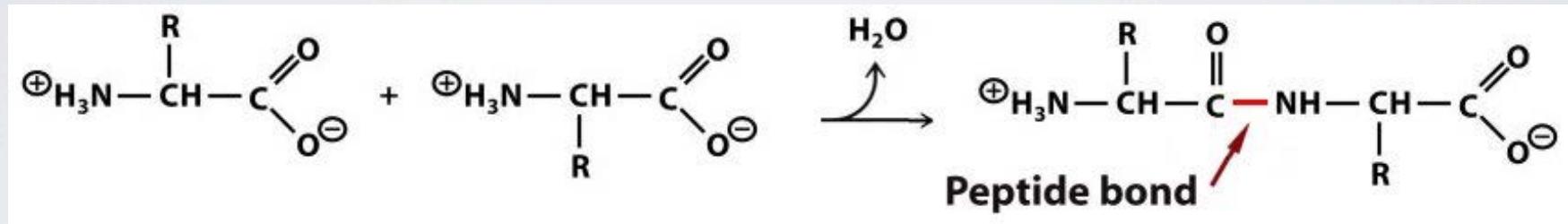


Image from: <http://www.ncbi.nlm.nih.gov/books/NBK21581/>

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

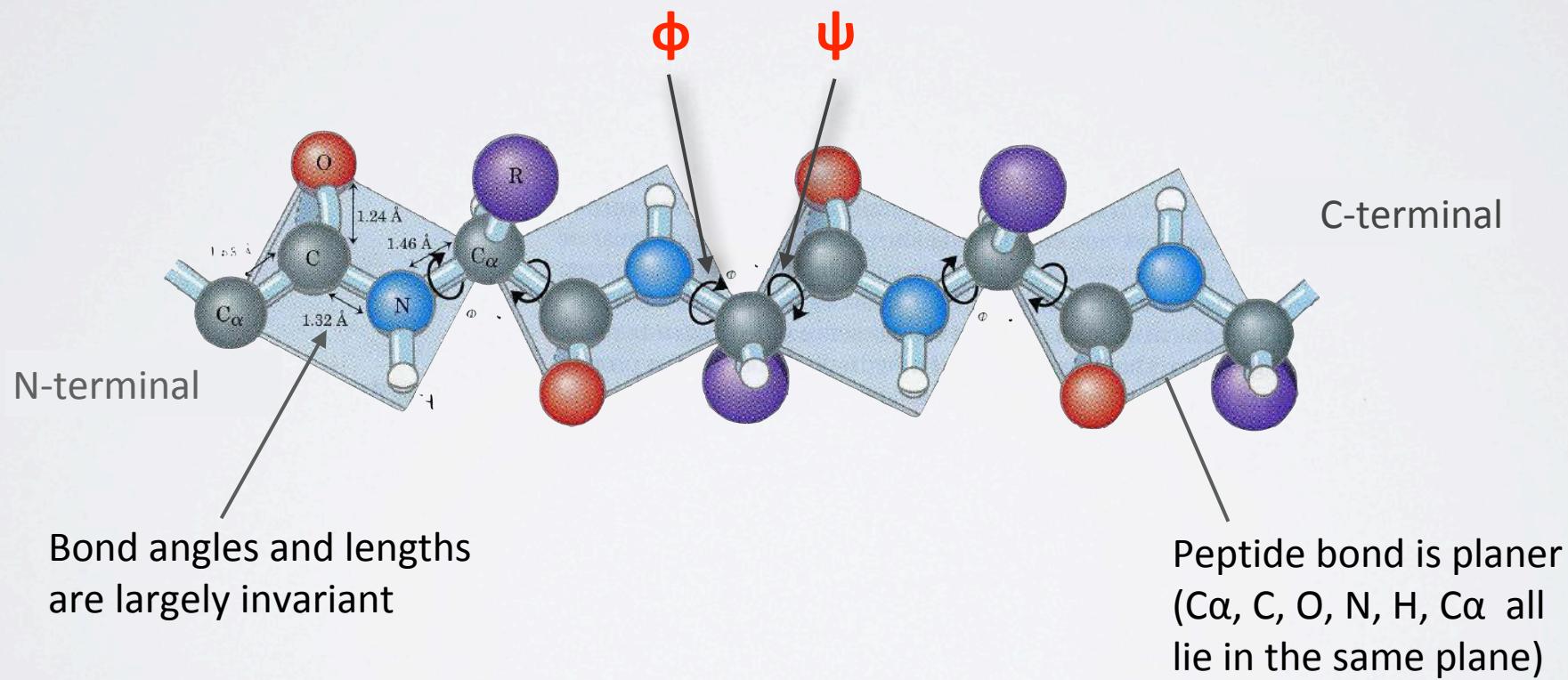
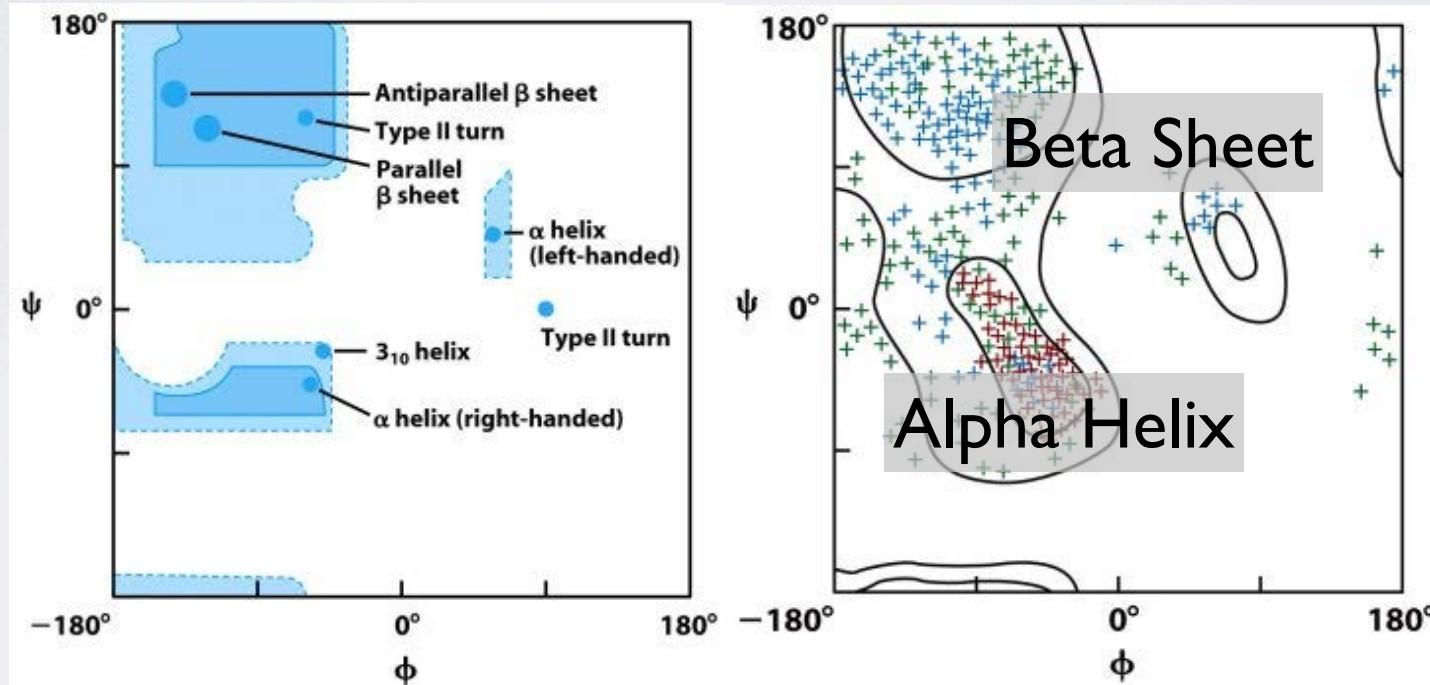


Image from: <http://www.ncbi.nlm.nih.gov/books/NBK21581/>

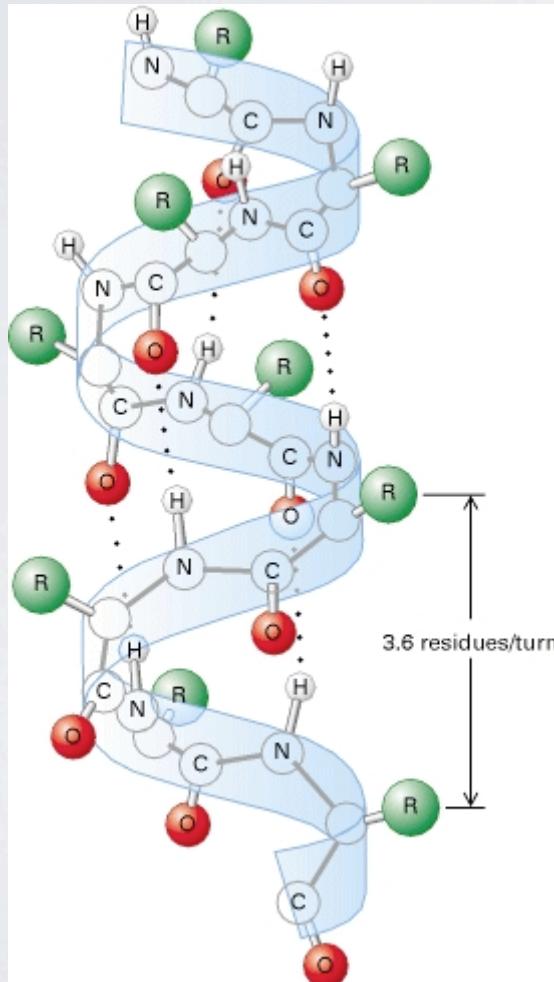
# PHI vs PSI PLOTS ARE KNOWN AS RAMACHANDRAN DIAGRAMS



- Steric hindrance dictates torsion angle preference
- Ramachandran plot show preferred regions of  $\phi$  and  $\psi$  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

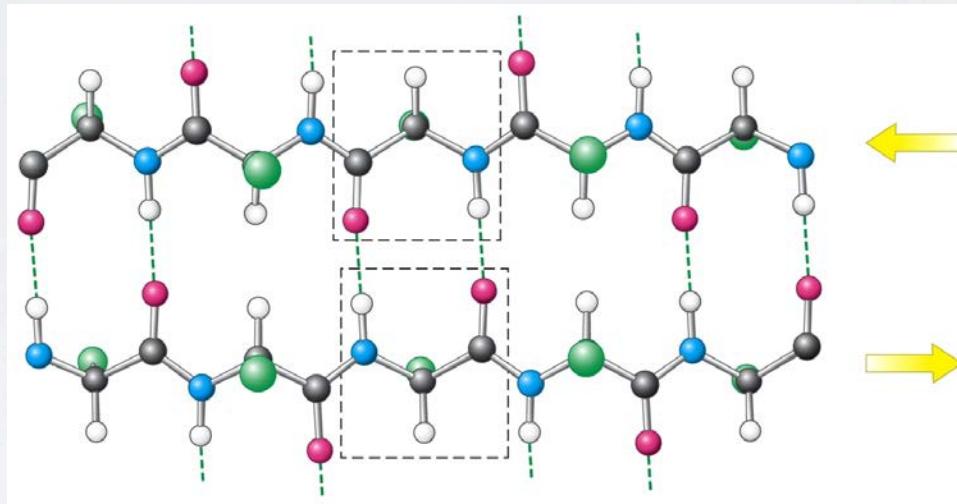


## $\alpha$ -helix $\beta$ -sheets

- Most common form has 3.6 residues per turn (number of residues in one full rotation of 360°)
- Hydrogen bonds (dashed lines) between residue  $i$  and  $i+4$  stabilize the structure
- The side chains (in green) protrude outward
- $3_{10}$ -helix and  $\pi$ -helix forms are less common

Hydrogen bond:  $i \rightarrow i+4$

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

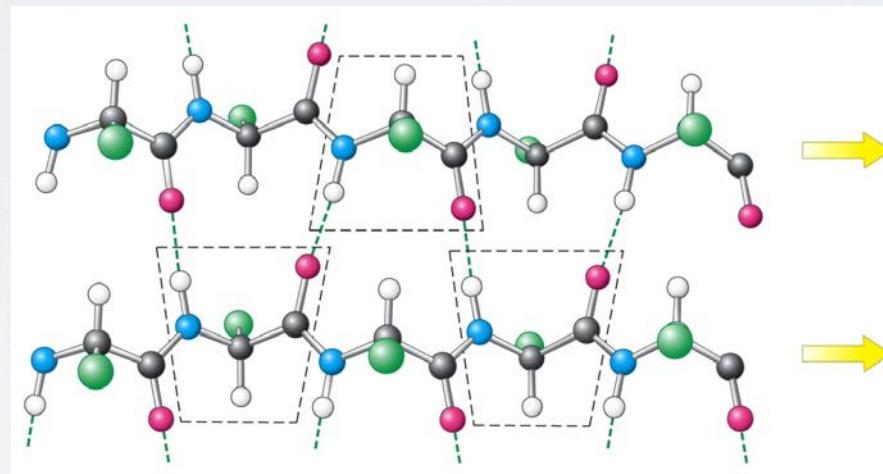


## In antiparallel $\beta$ -sheets

- Adjacent  $\beta$ -strands run in opposite 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**



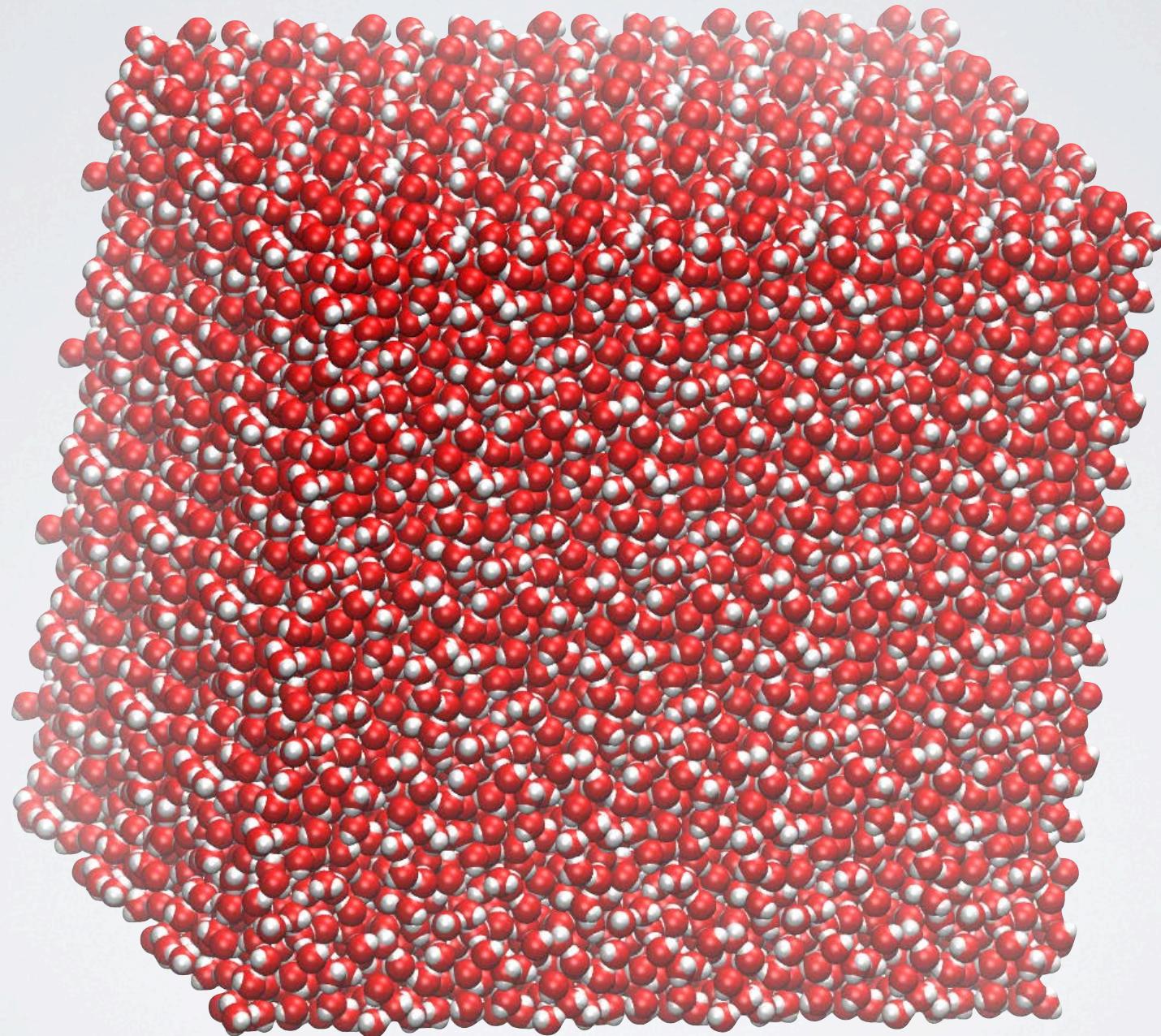
## In parallel $\beta$ -sheets

- Adjacent  $\beta$ -strands run in same direction
- Hydrogen bonds (dashed lines) between NH and CO stabilize the structure
- The side chains (in green) are above and below the sheet

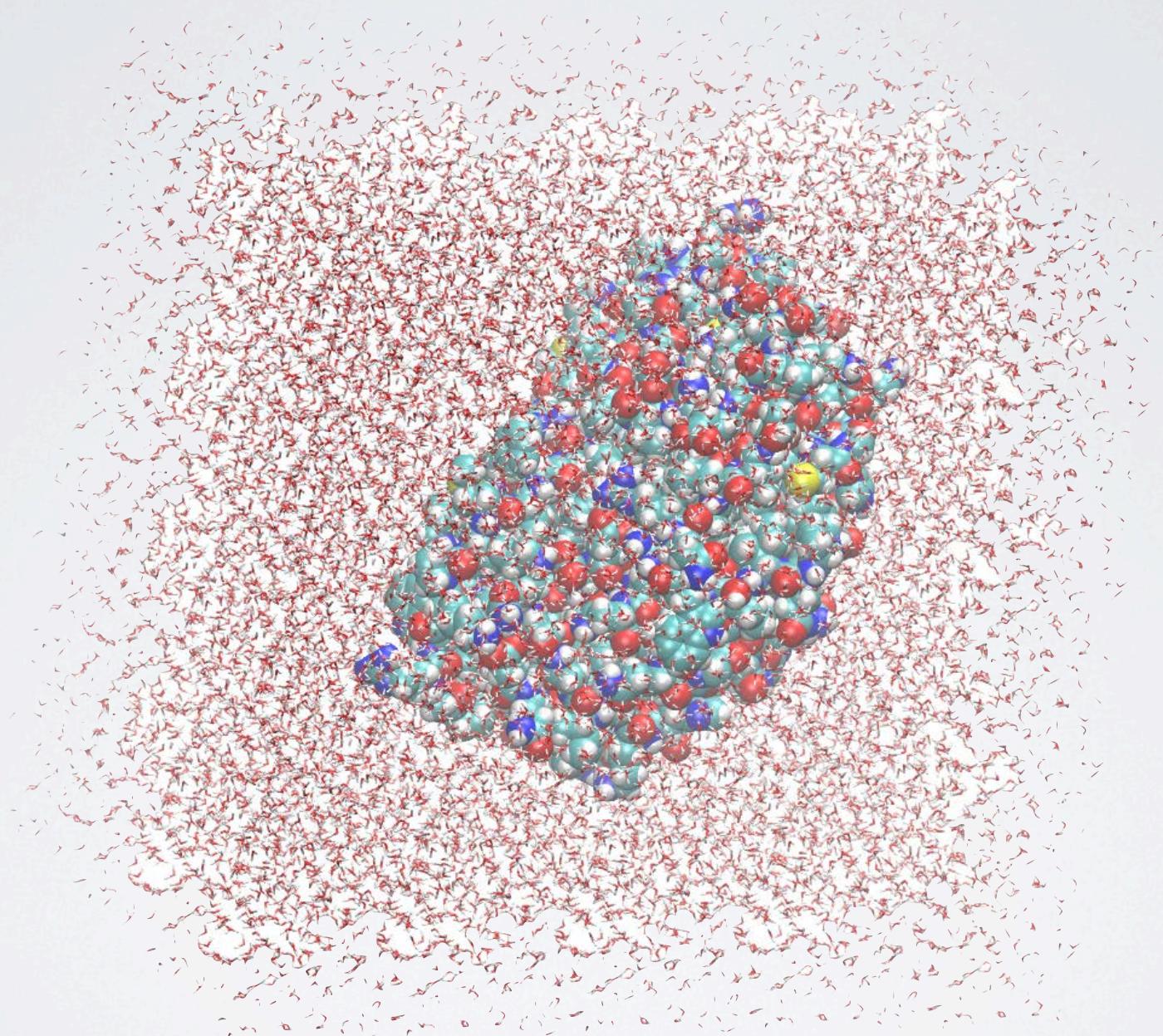
Image from: <http://www.ncbi.nlm.nih.gov/books/NBK21581/>

# WHAT DOES A PROTEIN LOOK LIKE?

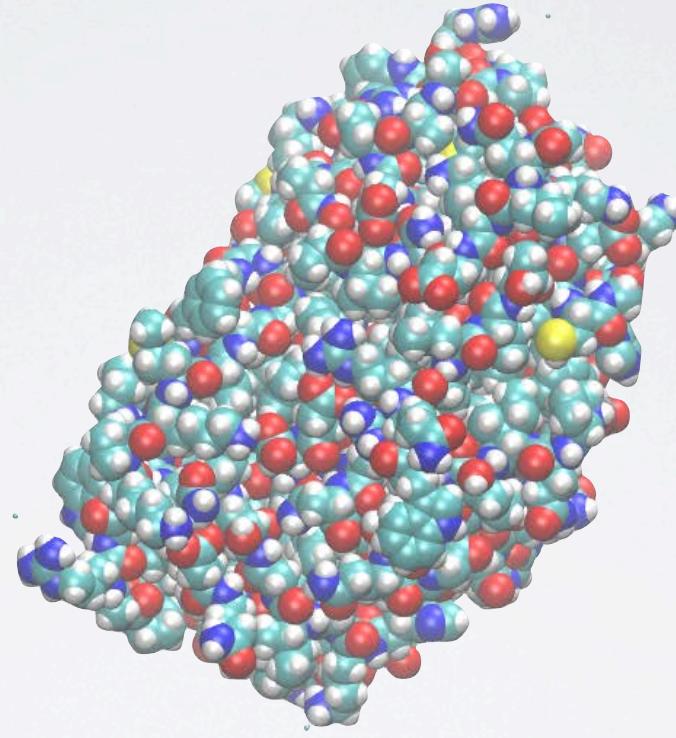
- Hidden in water?
- A close-packed globular object?
- A chain of connected secondary structures?



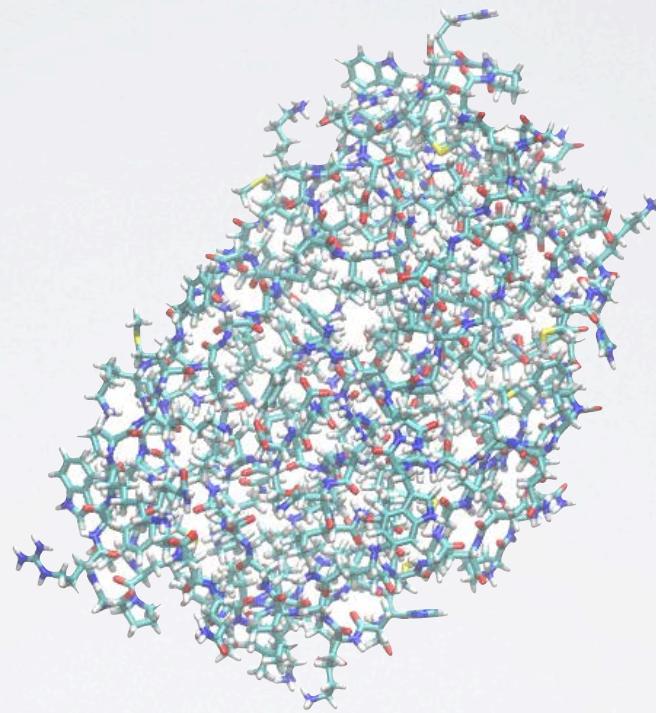
- Proteins are stable in water



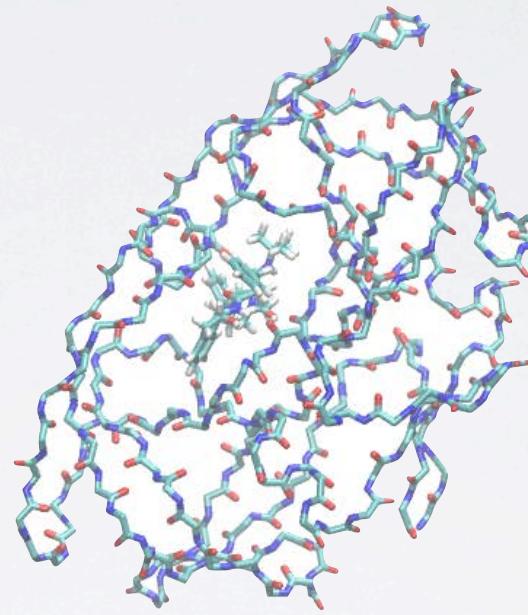
- Proteins closely interact with water



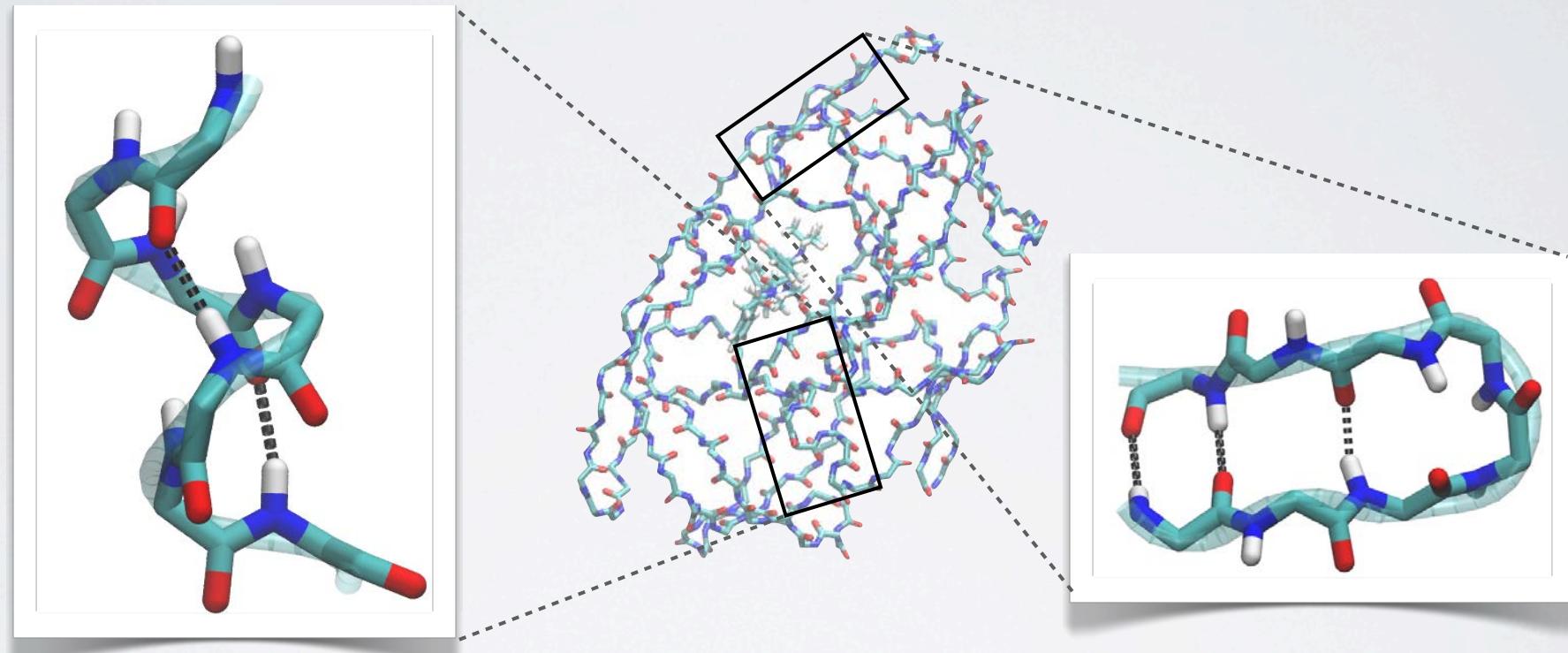
- Proteins are close packed solid but flexible objects



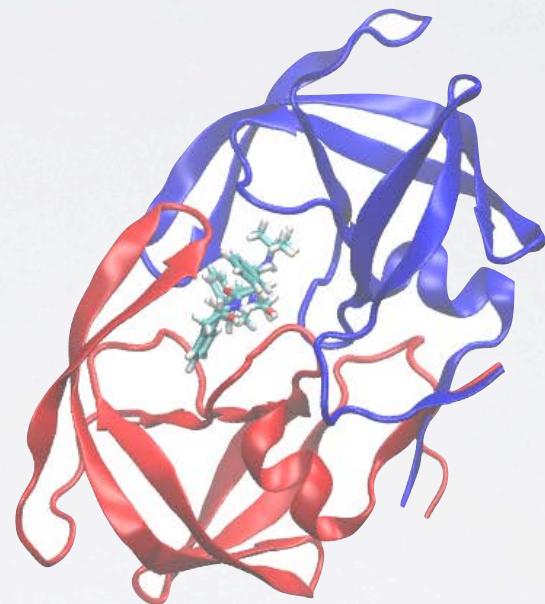
- Due to their large size and complexity it is often hard to see what's important in the structure



- Backbone or main-chain representation can help trace chain topology



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



- Simplified secondary structure representations are commonly used
- Now we can clearly see  $2^{\circ}$ ,  $3^{\circ}$  and  $4^{\circ}$  structure

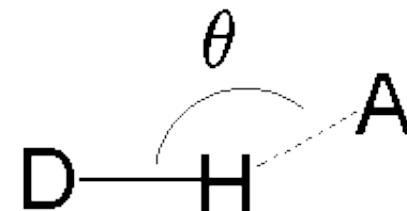
# Key forces affecting structure:

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

Hydrogen-bond donor      Hydrogen-bond acceptor



→ d →

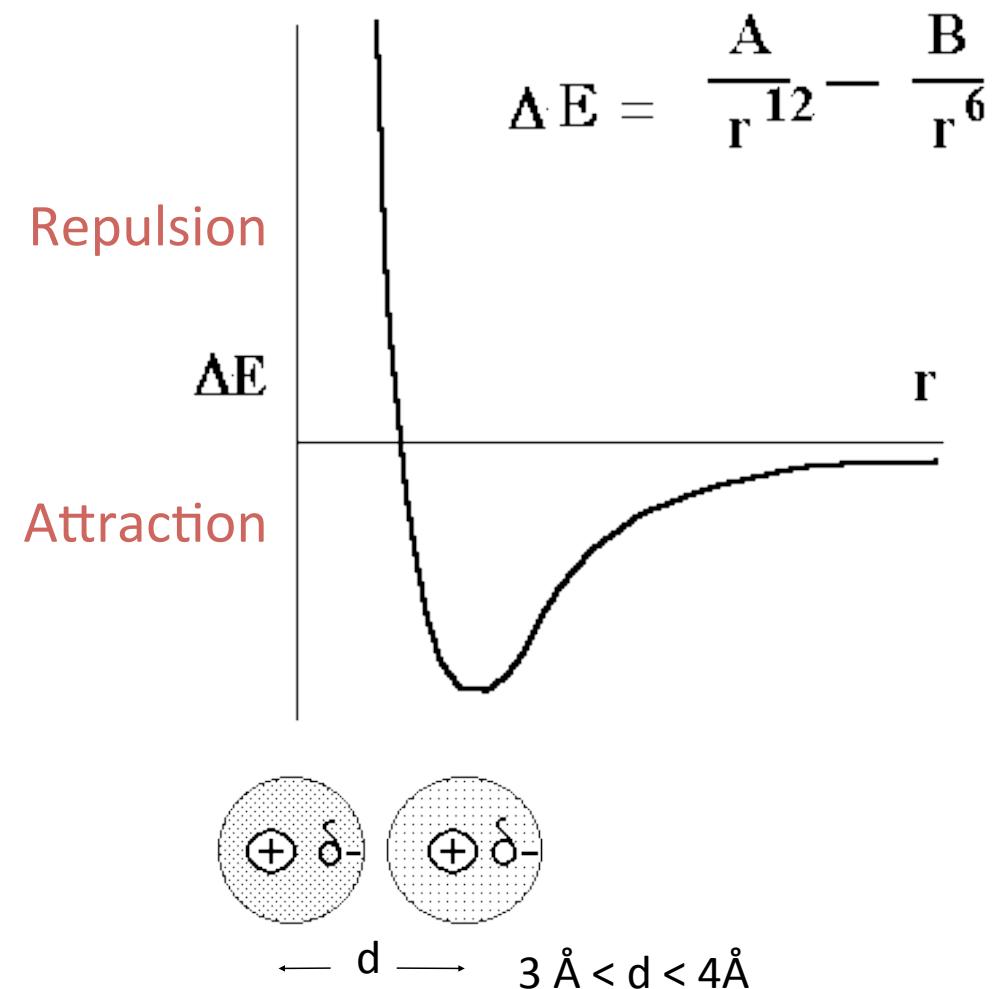


$2.6 \text{ \AA} < d < 3.1 \text{ \AA}$

$150^\circ < \theta < 180^\circ$

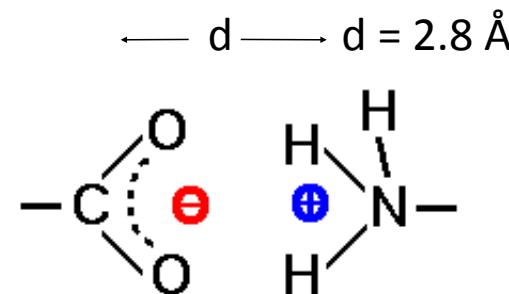
# Key forces affecting structure:

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



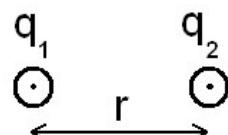
# Key forces affecting structure:

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



carboxyl group and amino group

(some time called IONIC BONDs or SALT BRIDGEs)



Coulomb's law

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

E = Energy

k = constant

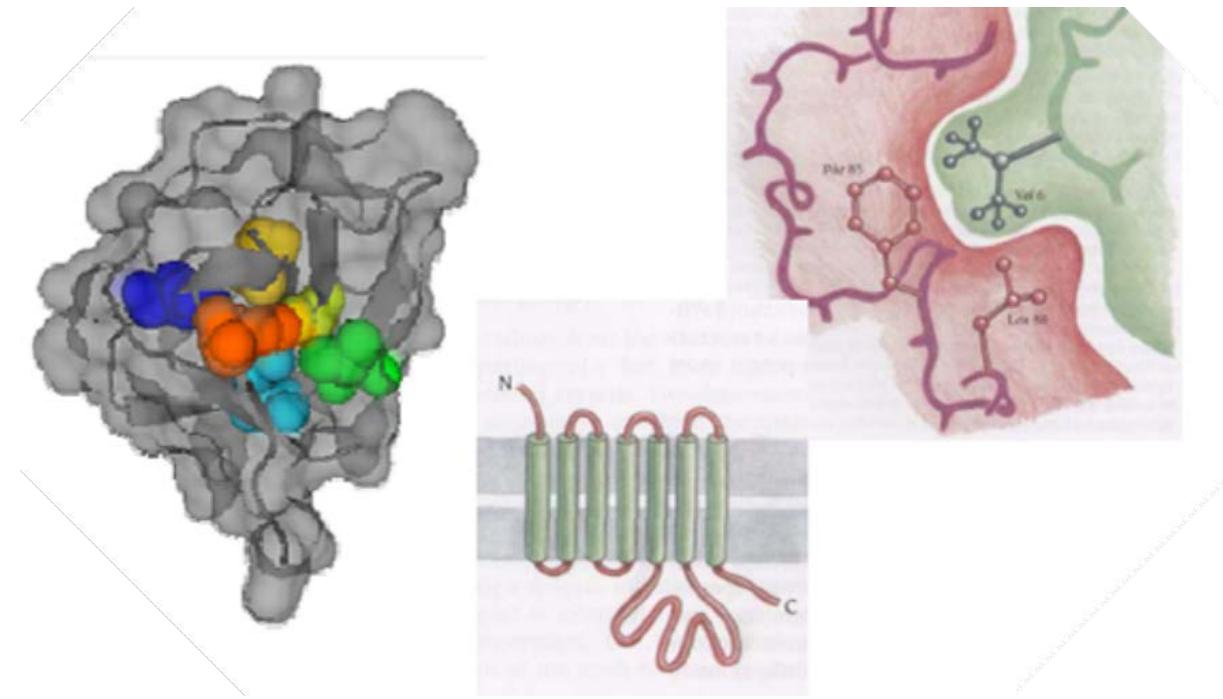
D = Dielectric constant (vacuum = 1; H<sub>2</sub>O = 80)

q<sub>1</sub> & q<sub>2</sub> = electronic charges (Coulombs)

r = distance (Å)

# Key forces affecting structure:

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

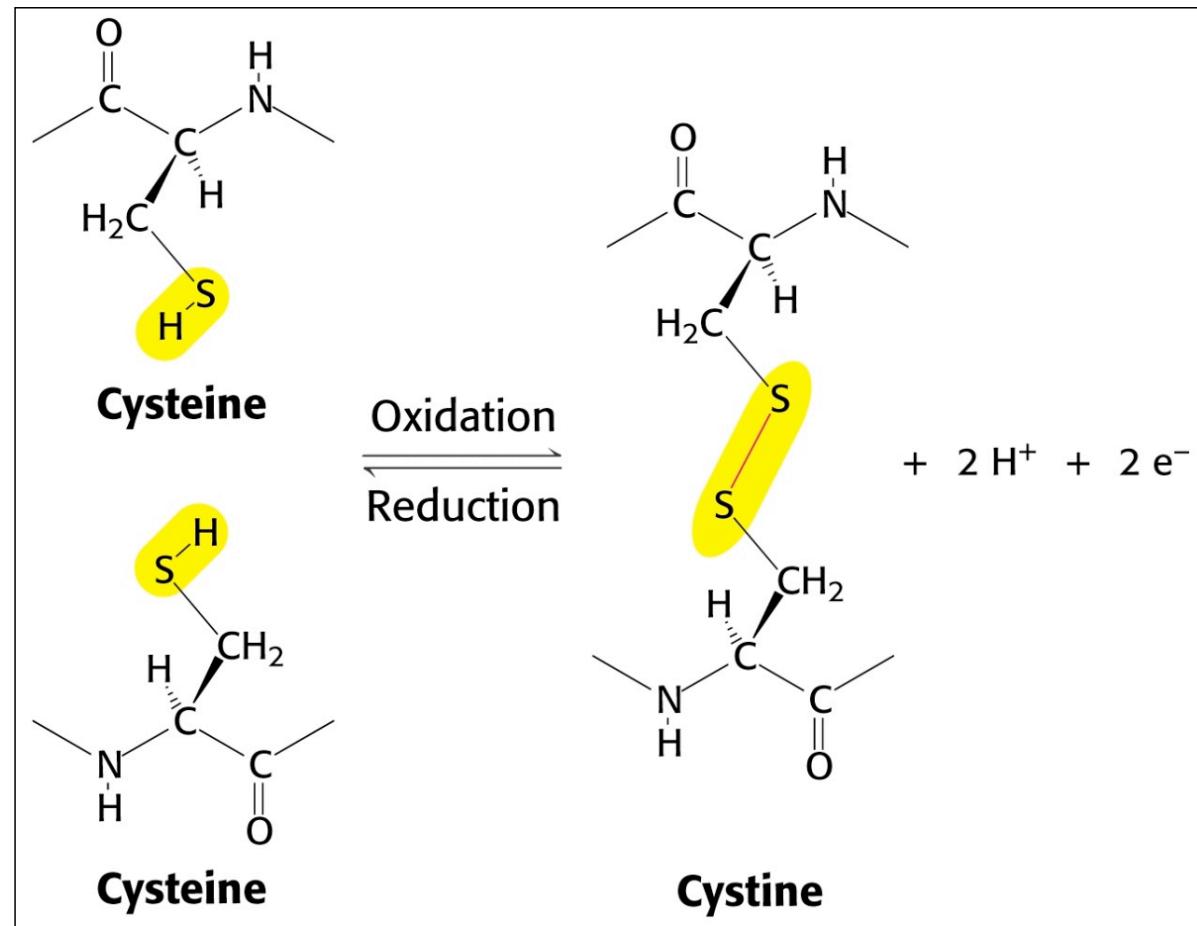


The force that causes hydrophobic molecules or nonpolar portions of molecules to aggregate together rather than to dissolve in water is called Hydrophobicity (Greek, “water fearing”). This is not a separate bonding force; rather, it is the result of the energy required to insert a nonpolar molecule into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

# Forces affecting structure:

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

Other names:  
cystine bridge  
disulfide bridge



Hair contains lots of disulfide bonds  
which are broken and reformed by heat

BREAK

# TODAY'S MENU:

- **Overview of structural bioinformatics**
  - Motivations, Goals and Challenges
- **Fundamentals of protein structure**
  - Structure composition, form and forces
- **Representing and interpreting biomolecular structure**
  - PDB and SCOP databases
  - Modeling energy as a function of structure
  - Physics based and knowledge based approaches
- **Example Application Areas**
  - Structure based drug discovery
    - Receptor and ligand based approaches
  - Predicting functional dynamics
    - Molecular dynamics and normal mode analysis
  - Protein structure and function prediction

RCSB Protein Data Bank - www.rcsb.org/pdb/home/home.do

Apps Home Docs Gmail Google Calendar Access 2CiteULike 2Papers 2Delicious Journals News&Sport Research Developments Other Bookmarks

**PDB** PROTEIN DATA BANK **PDB-101**

A MEMBER OF THE **PDB** | **EMDataBank**  
An Information Portal to Biological Macromolecular Structures  
As of Tuesday Nov 12, 2013 at 4 PM PST there are 95475 Structures | PDB Statistics |

Search Advanced Browse

Everything Author Macromolecule Sequence Ligand ?  
e.g., PDB ID, molecule name, author

Search History, Previous Results

**Biological Macromolecular Resource**

**Customize This Page**

**Available on the App Store**

**PDB-101** Hide Structural View of Biology **PDB-101** Hide Molecule of the Month **SNARE Proteins** Small membrane-enclosed vesicles are used like cargo trucks to deliver proteins and other molecules inside and outside of cells. When these vesicles reach their proper destination, they fuse with a membrane and deliver their cargo. For instance, vesicles are used inside cells to transport digestive enzymes from the Golgi to their final location in lysosomes. They are also used to deliver molecules out of the cell: for example, neurotransmitters are released from vesicles that fuse with the cell membrane at nerve synapses. The 2013 Nobel Prize was awarded to three researchers who have revealed the central molecular machinery for this process of vesicle fusion.

**MyPDB** Hide Login to your Account Register a New Account MyPDB Help Page

**Home** Hide News & Publications Usage/Reference Policies Deposition Policies Website FAQ Deposition FAQ Contact Us About Us Careers External Links Sitemap New Website Features

**Deposition** Hide All Deposit Services Electron Microscopy X-ray | NMR Validation Server BioSync Beamlines/Facilities Related Tools

**Tools** Hide Download Files Compare Structures Drug & Drug Target Mapping File Formats RESTful Web Services Widgets

**Help** Hide Launch Help System Display Settings Video Tutorials

**Explore Archive**

Organism Taxonomy  
Exp. Method X-ray Resolution  
Release Date Polymer Type  
Enzyme Classification SCOP Classification  
Protein Symmetry Protein Stoichiometry

**Experimental Method**

- X-ray (84373)
- Solution NMR (10116)
- Electron Microscopy (690)
- Solid-State NMR (62)
- Hybrid (59)
- Neutron Diffraction (43)
- Fiber Diffraction (38)
- Electron Crystallography (38)
- Solution Scattering (32)
- Other (24)

Show all

**Latest Structures**

4NCO : Crystal Structure of the BG505 SOSIP gp140 HIV-1 Env trimer in

**New Features** Hide **Latest release: September 2013** **Protein Symmetry and Stoichiometry** Visualize, browse & search symmetry / stoichiometry [Website Release Archive](#)

**RCSB PDB News** Hide Weekly | Quarterly | Yearly 2013-11-12 Upcoming Meeting: ABRCMS **2013 ABRCMS** November 13-16 **Nashville** Meet the RCSB PDB at booth 715 at the Annual Biomedical Research Conference for Minority Students (Nov 13-16; Nashville, TN). [more](#)

- Comparison Tool for Exploring Sequence and Structure Alignments
- Find Structures Using the Protein Symmetry Browser
- Fall Newsletter Published

**New Structures** Hide Latest Release New Structure Papers

**Search: HIV**

RCSB PDB - Query Results | www.rcsb.org/pdb/results/results.do?qrId=88EC9330&tabToShow=Current

A MEMBER OF THE **PDB** | **EMDataBank**  
An Information Portal to Biological Macromolecular Structures  
Nov 12, 2013 at 4 PM PST there are 95475 Structures | PDB Statistics | Other Bookmarks

**PDB-101**

Search Advanced Browse

Everything Author Molecule

HIV

Search History (1), Previous Results

2054 Structure Hits 109 Unreleased Structures 861 Citations 760 Ligand Hits 80 Web Page Hits

Query Details | Save Query to MyPDB

**Query Parameters:**  
Text Search for: hiv

**Other search suggestions:**

Molecule Name	Structural Domains	Molecule of the Month	Organism	Enzyme Classification
<ul style="list-style-type: none"> <li>HIV-1 protease ... (448)</li> <li>HIV-1 REVERSE ... (210)</li> <li>HIV-2 PROTEASE ... (18)</li> <li>Anti-HIV-1 reverse ... (6)</li> <li>HIV-1 fusion ... (4)</li> <li>HIV-1 DIS RNA (4)</li> </ul> <a href="#">More</a>	<ul style="list-style-type: none"> <li>HIV-1 reverse ... (112)</li> <li>HIV Type ... (180)</li> <li>HIV RNase ... (86)</li> <li>HIV-1 Transactivator ... (5)</li> <li>HIV-1 gp41 ... (2)</li> <li>HIV-1 Reverse ... (6)</li> </ul> <a href="#">More</a>	<ul style="list-style-type: none"> <li>Integrase [HIV]</li> <li>HIV-1 Protease</li> <li>HIV Capsid</li> <li>Reverse Transcriptase [HIV]</li> <li>T-Cell Receptor [HIV]</li> </ul>	<ul style="list-style-type: none"> <li>HIV-1 M:B_HXB2R (87)</li> <li>HIV-1 M:A (1)</li> <li>HIV-2 subtype A (9)</li> <li>HIV-1 M:J (1)</li> <li>HIV-1 M:J_SE9173 (1)</li> <li>HIV-1 M:B_ARV2/SF2 (52)</li> </ul> <a href="#">More</a>	<ul style="list-style-type: none"> <li>3.4.23.16: HIV-1 retropepsin (548)</li> <li>3.4.23.47: HIV-2 retropepsin (19)</li> </ul>

UniProt Gene Names	BIRD Molecules	Chemical Name	Ontology Terms	Pfam Description
<ul style="list-style-type: none"> <li>HIVEP1 (3)</li> <li>HIV1 ENV (4)</li> </ul> <a href="#">Find all</a>	<ul style="list-style-type: none"> <li>PRD_000280 - HIV entry ... (1)</li> <li>PRD_000281 - HIV ENTRY ... (2)</li> </ul> <a href="#">Find all</a>	<ul style="list-style-type: none"> <li>BE6: HIV-1 INHIBITOR</li> <li>BE5: HIV-1 INHIBITOR</li> </ul> <a href="#">Find all</a>	<ul style="list-style-type: none"> <li>HS : TAR (HIV-1) RNA ... (3)</li> <li>B04.820 ... 350: HIV [MeSH ... (1171)</li> <li>HS : TAR (HIV-1) RNA ... (3)</li> <li>D08.811 ... .187: HIV Reverse ... (218)</li> <li>HS : TAR (HIV-1) RNA ... (1)</li> <li>D27.505 ... Anti-HIV Agents ... (694)</li> </ul> <a href="#">More</a>	<ul style="list-style-type: none"> <li>PF13949 ... binding to HIV (10)</li> </ul>

**Query Refinements: Select an item or pie chart**

Organism	Taxonomy	Experimental Method	X-ray Resolution	Release Date
<ul style="list-style-type: none"> <li>Human immunodeficiency virus 1 (921)</li> <li>Hom sapiens (477)</li> <li>HIV-1 M:B_HXB2R (87)</li> <li>Mus musculus (85)</li> <li>Human immunodeficiency virus ty ... (60)</li> <li>Human immunodeficiency virus ty ... (58)</li> <li>HIV-1 M:B_ARV2/SF2 (52)</li> <li>Other (370)</li> </ul>	<ul style="list-style-type: none"> <li>Viruses (1464)</li> <li>Eukaryota (651)</li> <li>Unassigned (132)</li> <li>Bacteria (75)</li> <li>Other (25)</li> <li>Archaea (5)</li> </ul>	<ul style="list-style-type: none"> <li>X-ray (1735)</li> <li>Solution NMR (255)</li> <li>Electron Microscopy (56)</li> <li>Solid-State NMR (3)</li> <li>Other (2)</li> <li>Electron Crystallography (1)</li> <li>Neutron Diffraction (1)</li> <li>Hybrid (1)</li> </ul>	<ul style="list-style-type: none"> <li>less than 1.5 Å (183)</li> <li>1.5 - 2.0 Å (579)</li> <li>2.0 - 2.5 Å (496)</li> <li>2.5 - 3.0 Å (347)</li> <li>3.0 and more Å (131)</li> <li>more choices...</li> </ul>	<ul style="list-style-type: none"> <li>before 2000 (315)</li> <li>2000 - 2005 (372)</li> <li>2005 - 2010 (571)</li> <li>2010 - today (796)</li> <li>this year (239)</li> <li>this month (9)</li> <li>more choices...</li> </ul>

Polymer Type	Enzyme Classification	SCOP Classification	Protein Symmetry	Protein Stoichiometry
<ul style="list-style-type: none"> <li>Protein (1885)</li> <li>RNA (79)</li> <li>Mixed (76)</li> <li>DNA (14)</li> </ul>	<ul style="list-style-type: none"> <li>3: Hydrolases (841)</li> <li>2: Transferases (304)</li> <li>5: Isomerasers (26)</li> <li>6: Ligases (5)</li> <li>1: Oxidoreductases (1)</li> </ul>	<ul style="list-style-type: none"> <li>All beta proteins (529)</li> <li>Alpha and beta proteins (a/b) (128)</li> <li>Alpha and beta proteins (a+b) (119)</li> <li>Multi-domain proteins (alpha an ... (113)</li> <li>All alpha proteins (527)</li> </ul>	<ul style="list-style-type: none"> <li>Cyclic (913)</li> <li>Asymmetric (869)</li> <li>Dihedral (14)</li> <li>Helical (4)</li> <li>Tetrahedral (1)</li> </ul>	<ul style="list-style-type: none"> <li>Homomer (917)</li> <li>Heteromer (588)</li> <li>Monomer (296)</li> <li>more choices...</li> </ul>

Search: 1HSG  
(PDB ID)

RCSB Protein Data Bank - www.rcsb.org/pdb/explore/explore.do?structureId=1HSG

A MEMBER OF THE PDB | EMDDataBank  
An Information Portal to Biological Macromolecular Structures  
12, 2013 at 4 PM PST there are 95475 Structures | PDB Statistics | Other Bookmarks

**PDB-101**

Search Advanced Browse

Everything Author Macrom

1HSG

Search History (1), Previous Results (2054)

**PDB-101** Hide

- Structural View of Biology
- Understanding PDB Data
- Molecule of the Month
- Educational Resources
- Author Profiles

**MyPDB** Hide

- Login to your Account
- Register a New Account
- MyPDB Help Page
- Query Results (2054)
- Query History (1)

**Home** Hide

- News & Publications
- Usage/Reference Policies
- Deposition Policies
- Website FAQ
- Deposition FAQ
- Contact Us
- About Us
- Careers
- External Links
- Sitemap
- New Website Features

**Deposition** Hide

- All Deposit Services
- Electron Microscopy
- X-ray | NMR
- Validation Server
- BioSync Beamlines/Facilities
- Related Tools

**Tools** Hide

- Download Files
- Compare Structures
- Drug & Drug Target Mapping
- File Formats
- RESTful Web Services
- Widgets

**Help** Hide

- Launch Help System
- Display Settings
- Video Tutorials
- Glossary of Terms
- RCSB PDB Mobile

**Summary** 3D View Sequence Annotations Seq. Similarity 3D Similarity Literature Biol. & Chem. Methods Geometry Links

**CRYSTAL STRUCTURE AT 1.9 ANGSTROMS RESOLUTION OF HUMAN IMMUNODEFICIENCY VIRUS (HIV) II PROTEASE COMPLEXED WITH L-735,524, AN ORALLY BIOAVAILABLE INHIBITOR OF THE HIV PROTEASES**

1HSG Display Files Download Files Share this Page

DOI:10.2210/pdb1hsg/pdb

**Primary Citation**

Crystal structure at 1.9-A resolution of human immunodeficiency virus (HIV) II protease complexed with L-735,524, an orally bioavailable inhibitor of the HIV proteases.

Chen, Z., Li, Y., Chen, E., Hall, D.L., Darke, P.L., Culberson, C., Shafer, J.A., Kuo, L.C.

Journal: (1994) J.Biol.Chem. 269: 26344-26348

PubMed: 7929352 Search Related Articles in PubMed

**PubMed Abstract:**

L-735,524 is a potent, orally bioavailable inhibitor of human immunodeficiency virus (HIV) protease currently in a Phase II clinical trial. We report here the three-dimensional structure of L-735,524 complexed to HIV-2 protease at 1.9-A resolution, as well as the structure of the native HIV-2 protease at 2.5-A resolution. The structure of HIV-2 protease is found to be essentially identical to that of HIV-1 protease. In the crystal lattice of the HIV-2 protease complexed with L-735,524, the inhibitor is chelated to the active site of the homodimeric enzyme in one orientation. This feature allows an unambiguous assignment of protein-ligand interactions from the electron density map. Both Fourier and difference Fourier maps reveal clearly the closure of the flap domains of the protease upon L-735,524 binding. Specific interactions between the enzyme and the inhibitor include the hydroxy group of the hydroxymaminopentane amide moiety of L-735,524 ligating to the carboxyl groups of the essential Asp-25 and Asp-25' enzymic residues and the amide oxygens of the inhibitor hydrogen bonding to the backbone amide nitrogen of Ile-50 and Ile-50' via an intervening water molecule. A second bridging water molecule is found between the amide nitrogen N2 of L-735,524 and the carboxyl oxygen of Asp-29'. Although other hydrogen bonds also add to binding, an equally significant contribution to affinity arises from hydrophobic interactions between the protease and the inhibitor throughout the pseudo-symmetric S1/S1', S2/S2', and S3/S3' regions of the enzyme. Except for its pyridine ring, all lipophilic moieties (t-butyl, indanyl, benzyl, and piperidyl) of L-735,524 are rigidly defined in the active site.

**Keywords:**  
Aspartic Acid Endopeptidases, Binding Sites, Crystallography, X-Ray, Drug Resistance, HIV Protease, HIV Protease Inhibitors, Indinavir, Pyridines

**Related Structures:**  
Primary Citation of: 1HSG 1HSH 1HSI

**Organizational Affiliation:**  
Department of Biological Chemistry, Merck Research Laboratories, West Point, Pennsylvania 19486.

Click on abstract words and keywords to add them to the search box.

**Biological Assembly**

3D View More Images...  
Symmetry: C2 view  
Stoichiometry: Homo 2-mer - A2  
Biological assembly 1 assigned by authors and generated by PISA (software)  
Downloadable viewers:  
Simple Viewer Protein Workshop  
Kiosk Viewer

**MyPDB Personal Annotations** Hide

To save personal annotations, please login to your MyPDB account.

**Deposition Summary** Hide

Authors: Chen, Z.  
Deposition: 1995-03-31  
Release: 1995-04-03

# MOLECULAR MACHINERY: A Tour of the Protein Data Bank

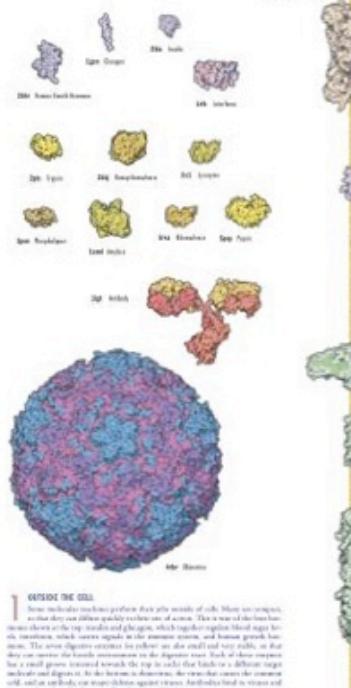
Living cells are filled with complex molecular machinery, a million times smaller than familiar machines like computers or automobiles. Cells use these tiny molecular machines to perform all of the jobs needed for life. Some are molecular switches that can load into cell-size pieces. Some have new roles when cells grow or when damaged tissues are repaired. Some are molecular lenses and muscles that support cells and help them move and crawl, fence fight off attackers, defending against infections.

Researchers around the world are studying these materials and determining their precise atomic structures. These structures are available on the Internet through the Protein Data Bank (<http://www.rcsb.org>), the central depository of biomolecular structures. At the top of the page, there is a search box with the following options:

- New**
- Recent**
- Most Cited**
- Supplemental PDB**

In these pictures, the Protein Data Bank is always a magnification of 1,880,140 Å. Since each structure is shown as a small square, many of them are arranged in several columns, which are indicated by different colors. An enormous range of sizes is shown here; the entire molecule at the left has only three atoms and the ribosome structure below has thousands.

By David S. Grotto, The Scripps Research Institute, La Jolla, California, 92093  
Photo credit: Ted R. Grotto, San Diego Environmental Art



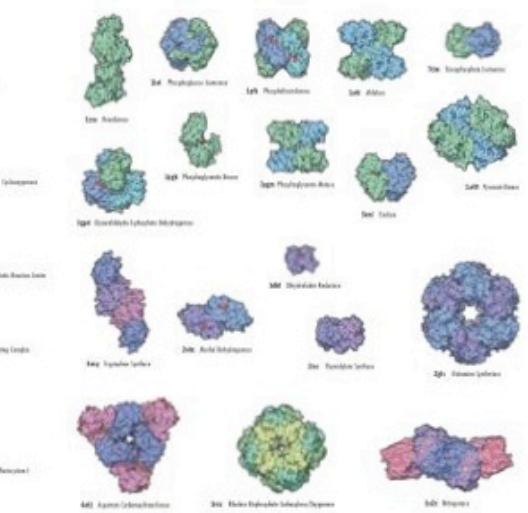
**1 OUTSIDE THE CELLS.** Cells have different problems than the inside of cells. When we eat sugar, most cells can utilize glucose exclusively as a source. But a few of the body have to use either fat or protein and glucose, which requires oxygen. Blood sugar is glucose, which serves as fuel in the immune system, and human brain cells. The active digestive enzymes for cellulose are the small and very stable, so that they do not break down the cellulose until it reaches the large intestine, where microorganisms break it down. The large intestine is where the large molecules and fibers are broken down. As the bacteria do their work, the bacteria release the enzymes and, as antibiotics, eat more cellulose against others. Antibiotics found in viruses will prevent these from breaking down the cellulose.

**MEMBRANES**  
 2. Cells are surrounded by a membrane made of lipids, like the phospholipid and cholesterol molecules shown at the top. Membranes have the cellular machinery needed and selected material. Many proteins are embedded in this membrane, performing a variety of essential tasks. ATP synthase is a protein that uses energy from the hydrolysis of ATP to move protons across the membrane. This creates a proton gradient that drives the rotation of the F1-ATPase subunit, which uses ATP synthase, and the protein conchoctase to move electrons from NADH. This is called redox in biochemistry. The small neural membrane molecule of a change shape takes information, changing the neurotransmitter and signal to a nerve cell. Endoplasmic reticulum is one of the main organelles in the cell. It is involved in protein folding, lipid synthesis, and calcium storage. It is also involved in the transport of proteins from one membrane involved in photosynthesis, which capture energy from light and use it to power the synthesis of sugar in plant cells.

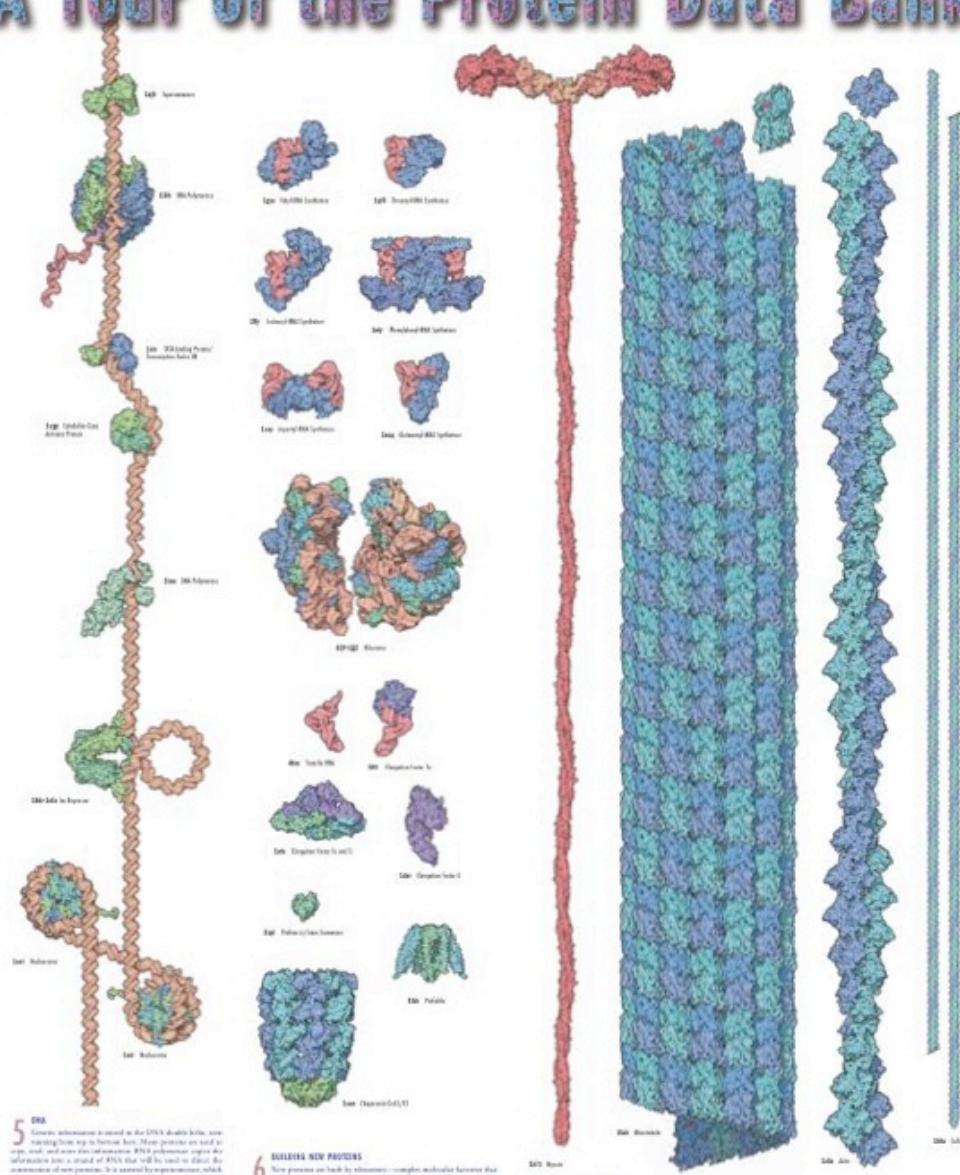
The diagram shows five distinct IgM antibody molecules, each with a unique Y-shape and color-coded regions:

- IgM Isotype:** A yellow molecule with four pink circular tips.
- IgM Proteasome:** A yellow molecule with four green circular tips.
- IgM Receptor-like:** A yellow molecule with four blue circular tips.
- IgM Variable:** A yellow molecule with four red circular tips.
- IgM FcγR:** A blue molecule with four purple circular tips.

**TRANSPORT AND STORAGE**  
**3.** *In contrast, a protein-based membrane would let in both ions and cells, because molecules could pass in and out more easily and get out. The box shows a membrane lacking fine-tun. Five proteins that form channels through the membrane are shown. To the right of the box are several half-filled proteins, including a receptor and various transporters. Homeostasis and migration carry charges. Homeostasis leaves a half-filled shell when ions pass in/out. Several arrows cause many different molecules in the blood.*



**Chemical Enzymes**  
Cells have a breeding variety of enzymes—proteins that perform chemical reactions. In the cap are the enzymes that perform glycolysis, the breakdown of sugar to form ATP. Below is a series of enzymes that perform  $\beta$ -oxidation reactions. Lipoproteins are membrane proteins that link membrane lipid and soluble hydrophobic fatty acids to soluble hydrophilic载体 proteins. The carrier protein is attached to the membrane via a hydrophobic tail. A carrier protein does not move across the cap. The carrier protein binds to the membrane via a hydrophobic tail. A carrier protein does not move across the membrane.



**5 RNA** Genetic information is stored in the DNA double helix, now we need to get that information out to produce proteins. To do this, we use RNA. RNA is a single strand of nucleic acid, and we can transcribe RNA from DNA. This process copies the information in a strand of DNA that will be used to direct the synthesis of a protein. The RNA strand is complementary to the DNA sequence when the two are transcribed, and goes through a series of modifications to become functional. RNA is also involved in DNA polymerase, which adds DNA nucleotides onto a DNA strand. Finally, the ribosomes are where the two processes link, because proteins are assembled on the ribosomes. The ribosomes are attached to the functional RNA strand.

**6 BUILDING NEW PROTEINS**

New proteins on hand by themselves complete molecular functions that the genetic code is about to direct construction. Many molecules

Slide Credit: RCSB PDB

### Structural Classification of Proteins



## Protein: Human immunodeficiency virus type 1 protease from Human immunodeficiency virus type 1 [TaxId: 11676]

SQ [P35963](#) 57-155 ! SQ [P04587](#) 69-167 ! SQ [P03366](#) 69-167 ! SQ [P03367](#) 69-167 ! SQ [P03368](#) 69-167

### Lineage:

1. Root: [scop](#)
2. Class: [All beta proteins](#) [48724]
3. Fold: [Acid proteases](#) [50629]  
*barrel, closed; n=6, S=10, complex topology*
4. Superfamily: [Acid proteases](#) [50630]  
*Superfamily*
5. Family: [Retroviral protease \(retropepsin\)](#) [50631]  
*dimer of identical mono-domain chains, each containing (6,10) barrel*
6. Protein: Human immunodeficiency virus type 1 protease [50632]
7. Species: [Human immunodeficiency virus type 1](#) [TaxId: 11676] [50633]  
SQ [P35963](#) 57-155 ! SQ [P04587](#) 69-167 ! SQ [P03366](#) 69-167 ! SQ [P03367](#) 69-167 ! SQ [P03368](#) 69-167

SCOP & CATH databases  
classify protein structural similarities

### PDB Entry Domains:

1. [2nmz](#)   
*automatically matched to d1s65a\_*  
*complexed with roc, so4; mutant*  
1. [region a:1-99](#) [138386]
2. [2nmz](#)   
*automatically matched to d1s65a\_*  
*complexed with roc, so4; mutant*  
1. [region b:101-199](#) [138387]
3. [3djk](#)   
*automatically matched to d1fgcc\_*  
*complexed with cl, g55, na; mutant*  
1. [region a:1-99](#) [157758]
4. [3djk](#)   
*automatically matched to d1fgcc\_*  
*complexed with cl, g55, na; mutant*

RCSB Protein Data Bank | SCOP: Protein: Human immunoglobulin | CATH: Protein Structure

www.cathdb.info

Apps Home Docs Gmail Google Calendar Access 2CiteULike 2Papers 2Delicious Journals News&Sport + Research Development Other Bookmarks

CATH Home Search Browse Download About Support Search CATH by keywords or ID

# CATH / Gene3D

16 million protein domains classified into 2,626 superfamilies

Browse » Search » Download » Take the Tour »

## What's New?

The CATH website has recently undergone a big overhaul. We really hope you find the new pages more useful, easier to use and quicker to load. Please [get in touch](#) and let us know what you think.

## Searching CATH

- Search by ID / keyword
- Search by FASTA sequence
- Search by PDB structure

## Example pages

- PDB "2bop"
- Domain "1cukA01"
- Relatives of "1cukA01"
- Superfamily "HUPs"
- Functional Family
- FunFam Alignment
- Search for "enolase"
- Superfamily Comparison

## Citing CATH

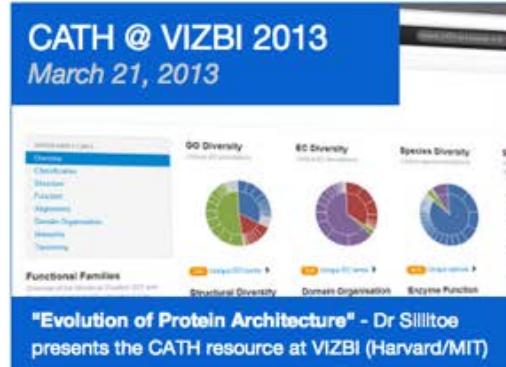
If you find this resource useful, please consider citing the reference that describes this work:

### New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures.

Sillitoe I, Cuff AL, Dessimoz BH, Dawson NL, Furnham N, Lee D, Lees JG, Lewis TE, Studer RA, Rentsch R, Yeats C, Thornton JM, Orengo CA

Nucleic Acids Res. 2013 Jan; 23203873

## Latest News



## Latest Release

**CATH v3.5** based on PDB dated September 20, 2011

173,536	CATH Domains
2,626	CATH Superfamilies
51,334	PDBs

**Gene3D v11** released March 18, 2012

1,639	Cellular Genomes
1,016	Viral Genomes
14,963,305	Protein Sequences
16,297,076	CATH Domain Predictions

### CATH News

- Support
- Jobs

### Get Started

- Documentation
- Tutorials

### Download

- WebServices
- Software

### About

- Orengo Group
- Web accessibility

RCSB Protein Data Bank | SCOP: Protein: Human integrin α<sub>1</sub> | CATH Superfamily 2.40.70.10

www.cathdb.info/version/latest/superfamily/2.40.70.10

Apps Home Docs Gmail Google Calendar Access 2CiteULike 2Papers 2Delicious Journals News&Sport + Research Development Other Bookmarks

CATH Home Search Browse Download About Support 1hsq

# CATH Superfamily 2.40.70.10

## Acid Proteases

Home / Superfamily 2.40.70.10

SUPERFAMILY LINKS

**Summary**

- Superfamily Superposition
- Classification / Domains
- Alignments
- Structural Neighbourhood
- Functional Annotations
- Taxonomy Browser
- Multi-Domain Organisation

**Functional Families**

Overview of the Structural Clusters (SC) and Functional Families (FF) within this CATH Superfamily

SC:1 Aspartic pn

GO Diversity

Unique GO annotations

111 Unique GO terms >

EC Diversity

Unique EC annotations

36 Unique EC terms >

Species Diversity

Unique species annotations

1468 Unique species >

Structures

Domains: 2031

Domains (< 95% seq id): 149

Domains (< 35% seq id): 30

Unique PDBs: 832

Alignments

Structural Clusters: 1

FunFam Clusters: 1

Function

Unique EC: 36

Unique GO: 111

Taxonomy

Unique Species: 1468

Sequence/Structure Diversity

Overview of the sequence / structure diversity of this superfamily compared to other superfamilies in CATH. Click

**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

# PHYSICS-BASED POTENTIALS

## ENERGY TERMS FROM PHYSICAL THEORY

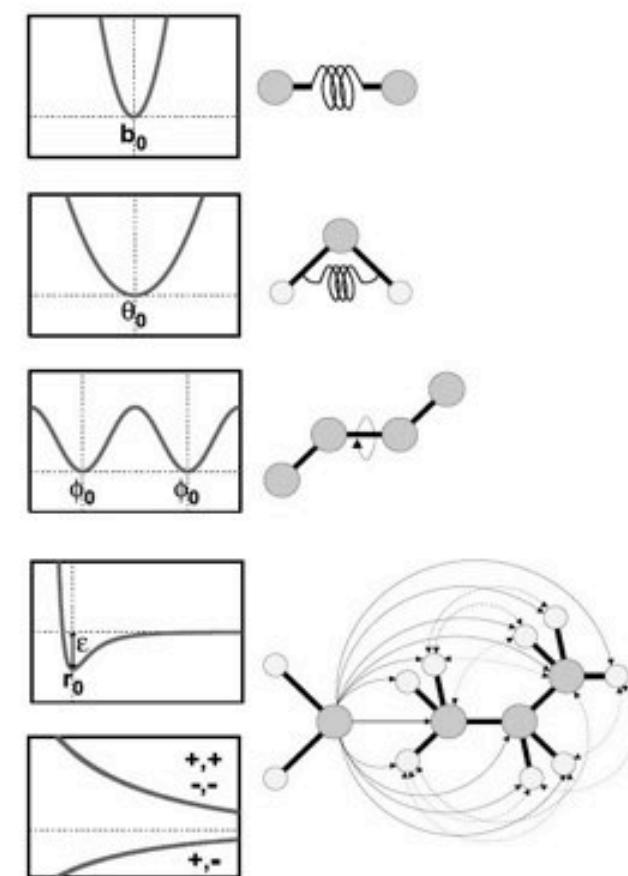
$$U(\vec{R}) = \underbrace{\sum_{bonds} k_i^{bond} (r_i - r_0)^2}_{U_{bond}} + \underbrace{\sum_{angles} k_i^{angle} (\theta_i - \theta_0)^2}_{U_{angle}} + \\ \underbrace{\sum_{dihedrals} k_i^{dih} [1 + \cos(n_i \phi_i + \delta_i)]}_{U_{dihedral}} + \\ \underbrace{\sum_i \sum_{j \neq i} 4\epsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \sum_i \sum_{j \neq i} \frac{q_i q_j}{\epsilon r_{ij}}}_{U_{nonbond}}$$

$U_{bond}$  = oscillations about the equilibrium bond length

$U_{angle}$  = oscillations of 3 atoms about an equilibrium bond angle

$U_{dihedral}$  = torsional rotation of 4 atoms about a central bond

$U_{nonbond}$  = non-bonded energy terms (electrostatics and Lenard-Jones)



# PHYSICS-ORIENTED APPROACHES

## Weaknesses

Fully physical detail becomes computationally intractable

Approximations are unavoidable

(Quantum effects approximated classically, water may be treated crudely)

Parameterization still required

## Strengths

Interpretable, provides guides to design

Broadly applicable, in principle at least

Clear pathways to improving accuracy

## Status

Useful, far from perfect

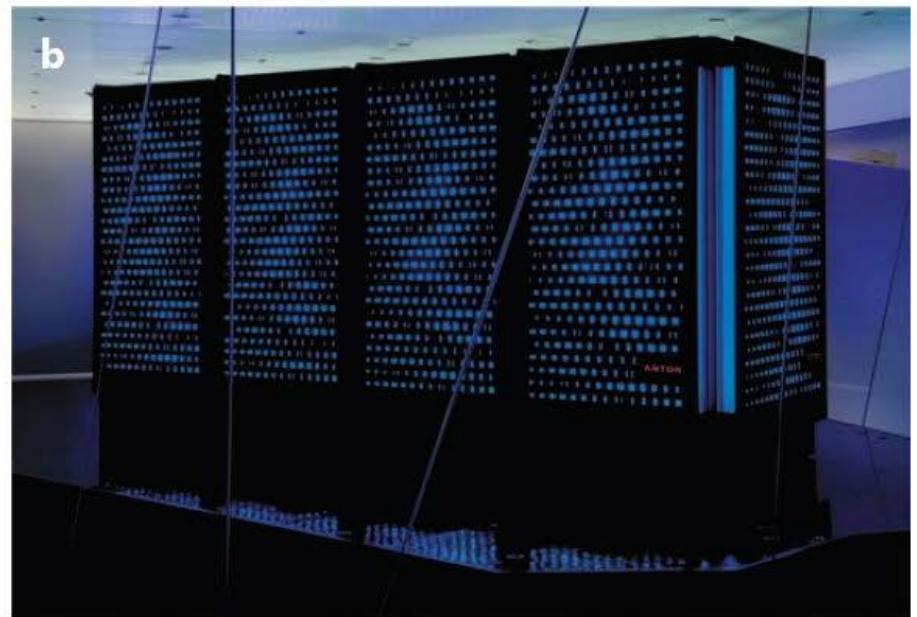
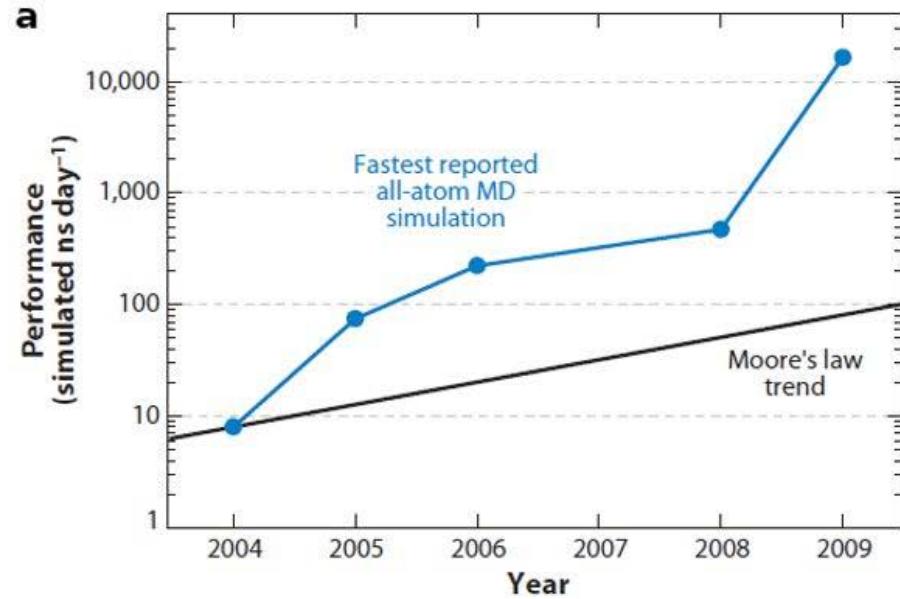
Multiple groups working on fewer, better approxs

Force fields, quantum

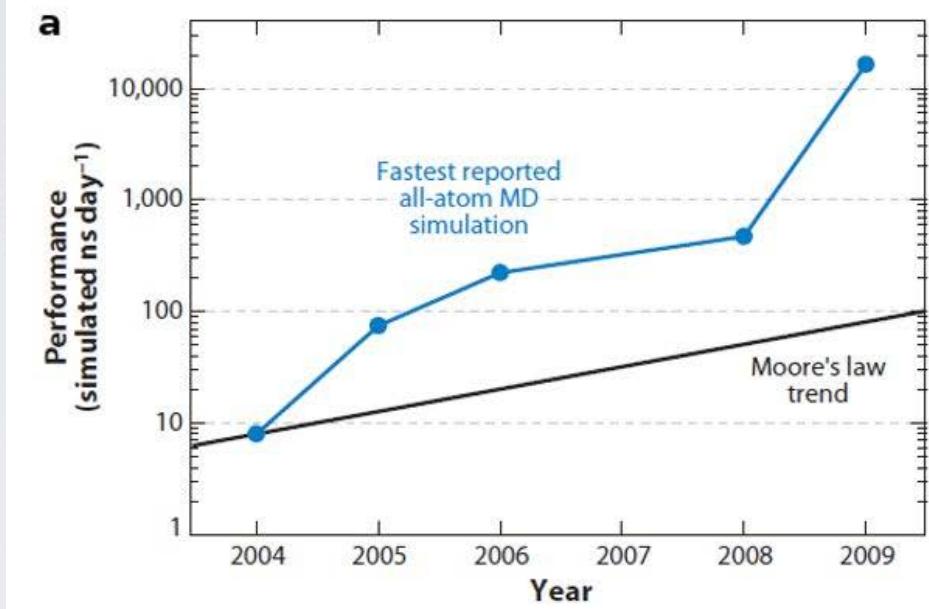
entropy, water effects

Moore's law: hardware improving

# SIDE-NOTE: GPUS AND ANTON SUPERCOMPUTER



## SIDE-NOTE:GPUS AND ANTON SUPERCOMPUTER

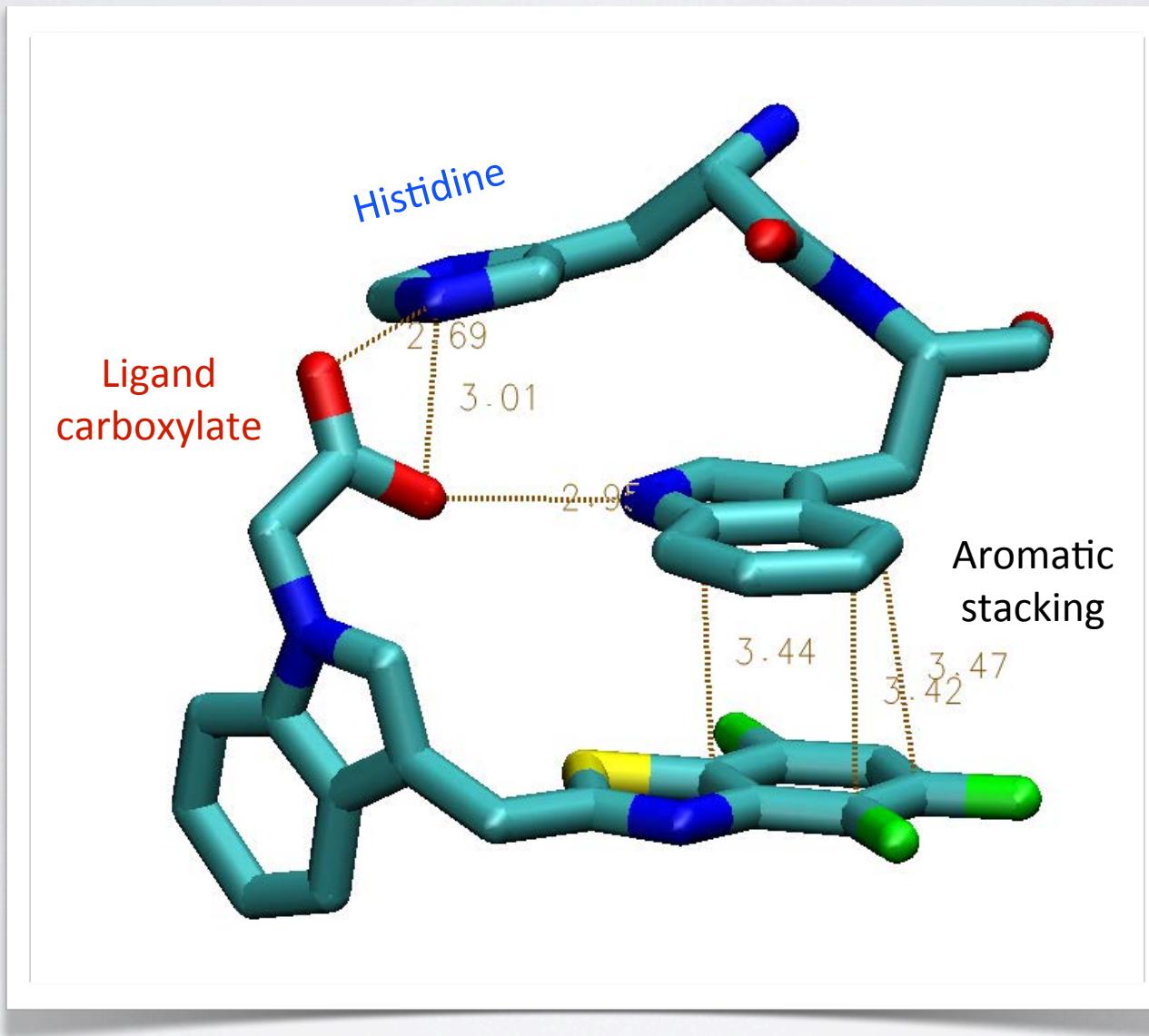


# **KEY CONCEPT:** POTENTIAL FUNCTIONS DESCRIBE A SYSTEMS **ENERGY** AS A FUNCTION OF ITS **STRUCTURE**

Two main approaches:

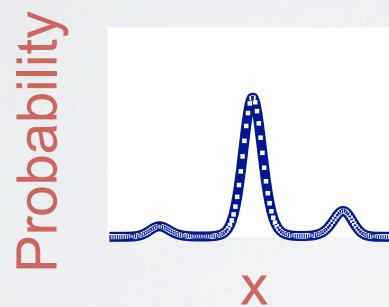
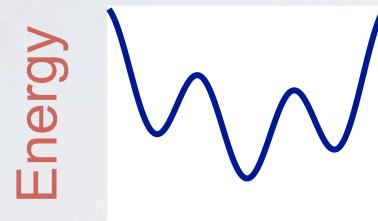
- (1). Physics-Based
- (2). Knowledge-Based

# KNOWLEDGE-BASED DOCKING POTENTIALS



# ENERGY DETERMINES **PROBABILITY** (STABILITY)

Basic idea: Use probability as a proxy for energy



Boltzmann:

$$p(r) \propto e^{-E(r)/RT}$$

Inverse Boltzmann:

$$E(r) = -RT \ln[p(r)]$$

Example: ligand carboxylate O to protein histidine N

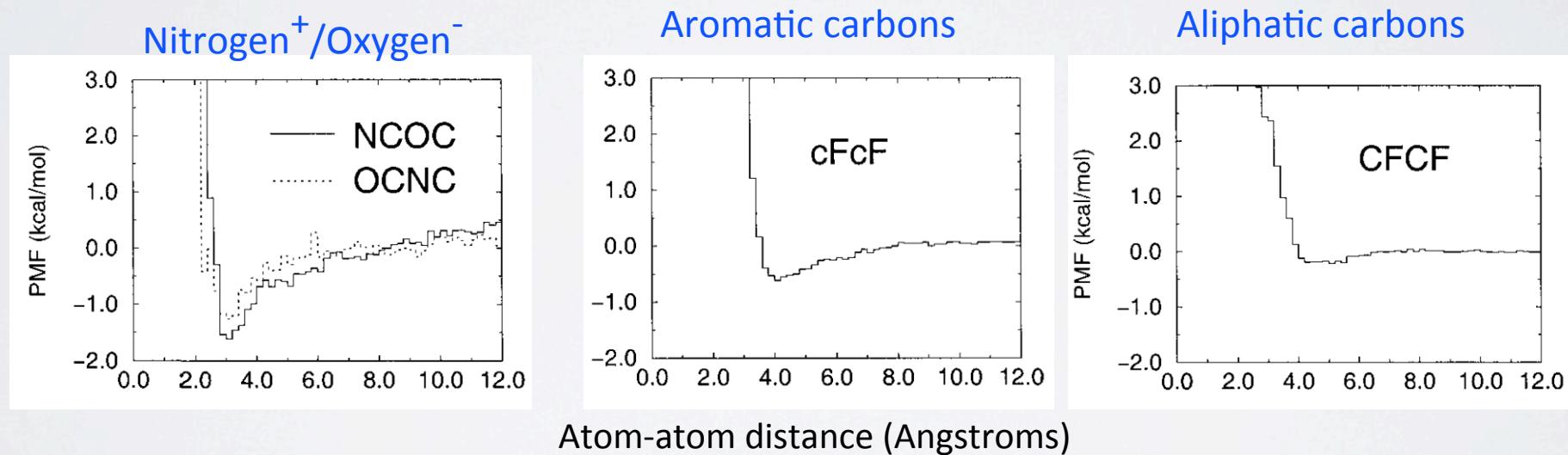
Find all protein-ligand structures in the PDB with a ligand carboxylate O

1. For each structure, histogram the distances from O to every histidine N
2. Sum the histograms over all structures to obtain  $p(r_{O-N})$
3. Compute  $E(r_{O-N})$  from  $p(r_{O-N})$

# KNOWLEDGE-BASED DOCKING POTENTIALS

“PMF”, Muegge & Martin, J. Med. Chem. (1999) 42:791

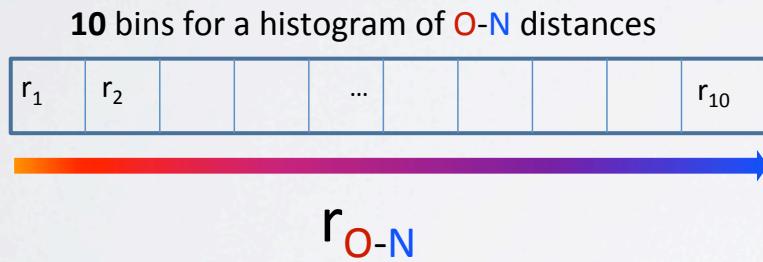
A few types of atom pairs, out of several hundred total



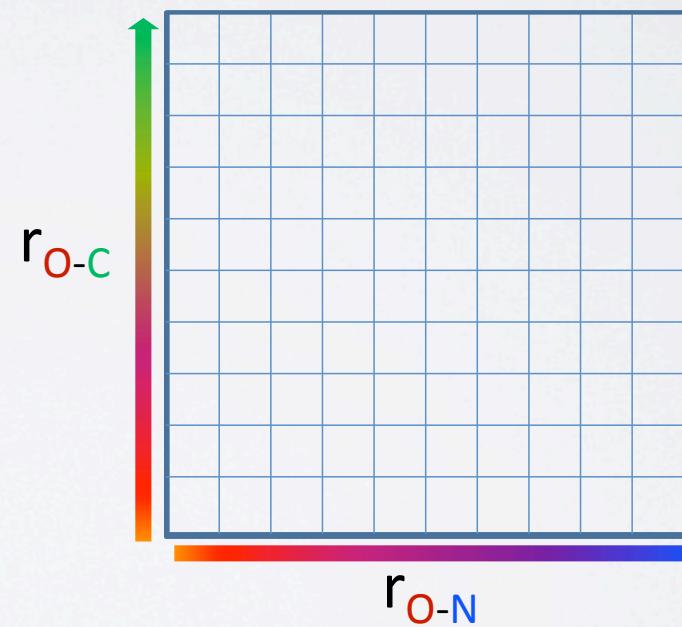
$$E_{prot-lig} = E_{vdw} + \sum_{pairs(ij)} E_{type(ij)}(r_{ij})$$

# LIMITATIONS OF KNOWLEDGE-BASED POTENTIALS

## 1. Statistical limitations (e.g., to pairwise potentials)



100 bins for a histogram of O-N & O-C distances



## 2. Even if we had infinite statistics, would the results be accurate? (Is inverse Boltzmann quite right? Where is entropy?)

# KNOWLEDGE-ORIENTED APPROACHES

## Weaknesses

Accuracy limited by availability of data

Accuracy may also be limited by overall approach

## Strengths

Relatively easy to implement

Computationally fast

## Status

Useful, far from perfect

May be at point of diminishing returns

(not always clear how to make improvements)

BREAK

# TODAY'S MENU:

- **Overview of structural bioinformatics**
  - Motivations, Goals and Challenges
- **Fundamentals of protein structure**
  - Structure composition, form and forces
- **Representing and interpreting biomolecular structure**
  - PDB and SCOP databases
  - Modeling energy as a function of structure
    - Physics based and knowledge based approaches
- **Example Application Areas**
  - Structure based drug discovery
    - Receptor and ligand based approaches
  - Predicting functional dynamics
    - Molecular dynamics and normal mode analysis
  - Protein structure and function prediction

# THE TRADITIONAL EMPIRICAL PATH TO DRUG DISCOVERY

**Compound library**  
(commercial, in-house,  
synthetic, natural)

**High throughput screening**  
(HTS)

**Hit confirmation**

**Lead compounds**  
(e.g.,  $\mu\text{M } K_d$ )

**Lead optimization**  
(Medicinal chemistry)

**Animal and clinical evaluation** ← **Potent drug candidates**  
( $\text{nM } K_d$ )

# COMPUTER-AIDED LIGAND DESIGN

Aims to reduce number of compounds synthesized and assayed

Lower costs

Reduce chemical waste

Facilitate faster progress

Two main approaches:

**(1). Receptor/Target-Based**

**(2). Ligand/Drug-Based**

Two main approaches:

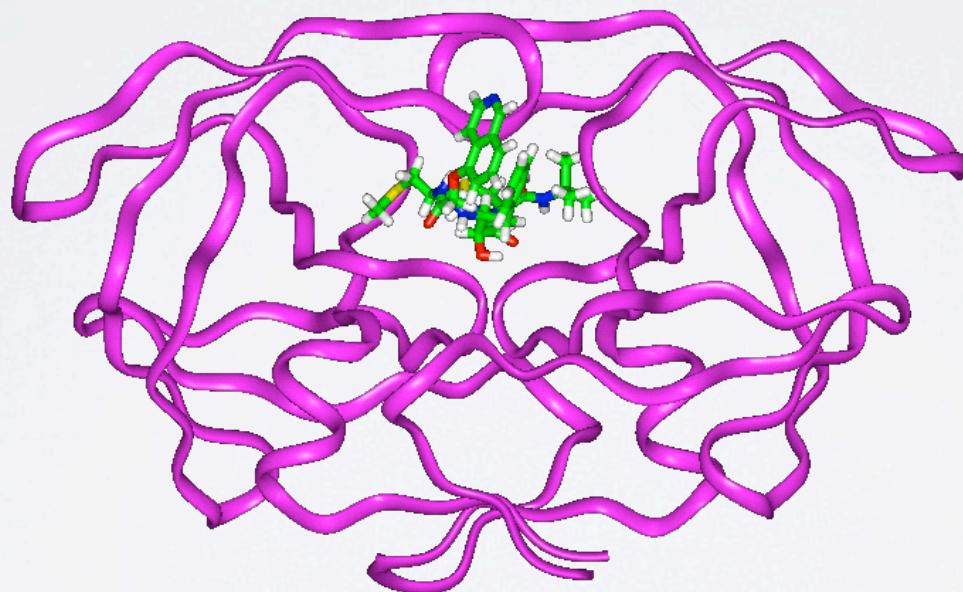
**(1). Receptor/Target-Based**

**(2). Ligand/Drug-Based**

# **SCENARIO I:**

## RECEPTOR-BASED DRUG DISCOVERY

Structure of Targeted Protein Known: **Structure-Based Drug Discovery**



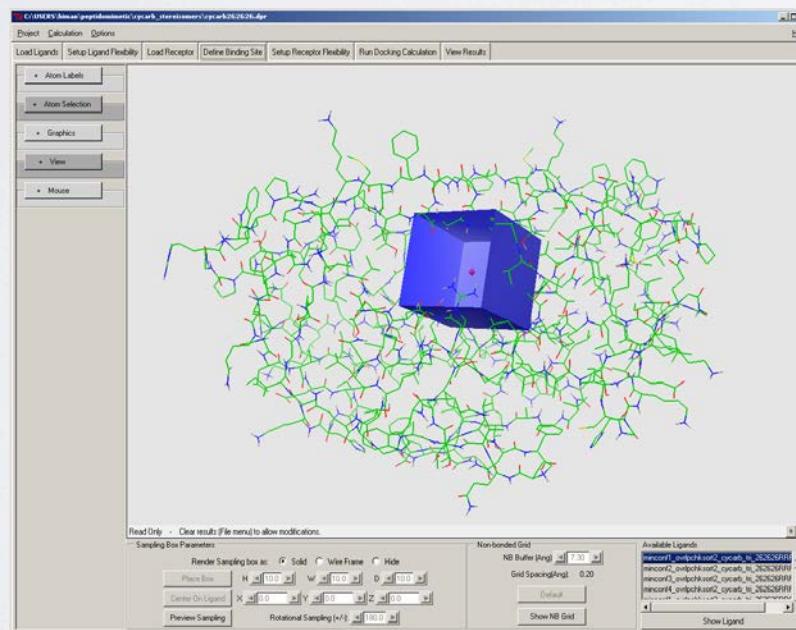
HIV Protease/KNI-272 complex

# PROTEIN-LIGAND DOCKING

# Structure-Based Ligand Design

## Docking software

## Search for structure of lowest energy



## Potential function

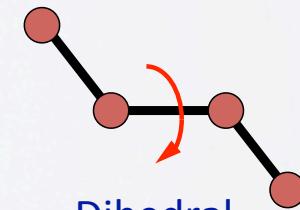
### Energy as function of structure



VDW

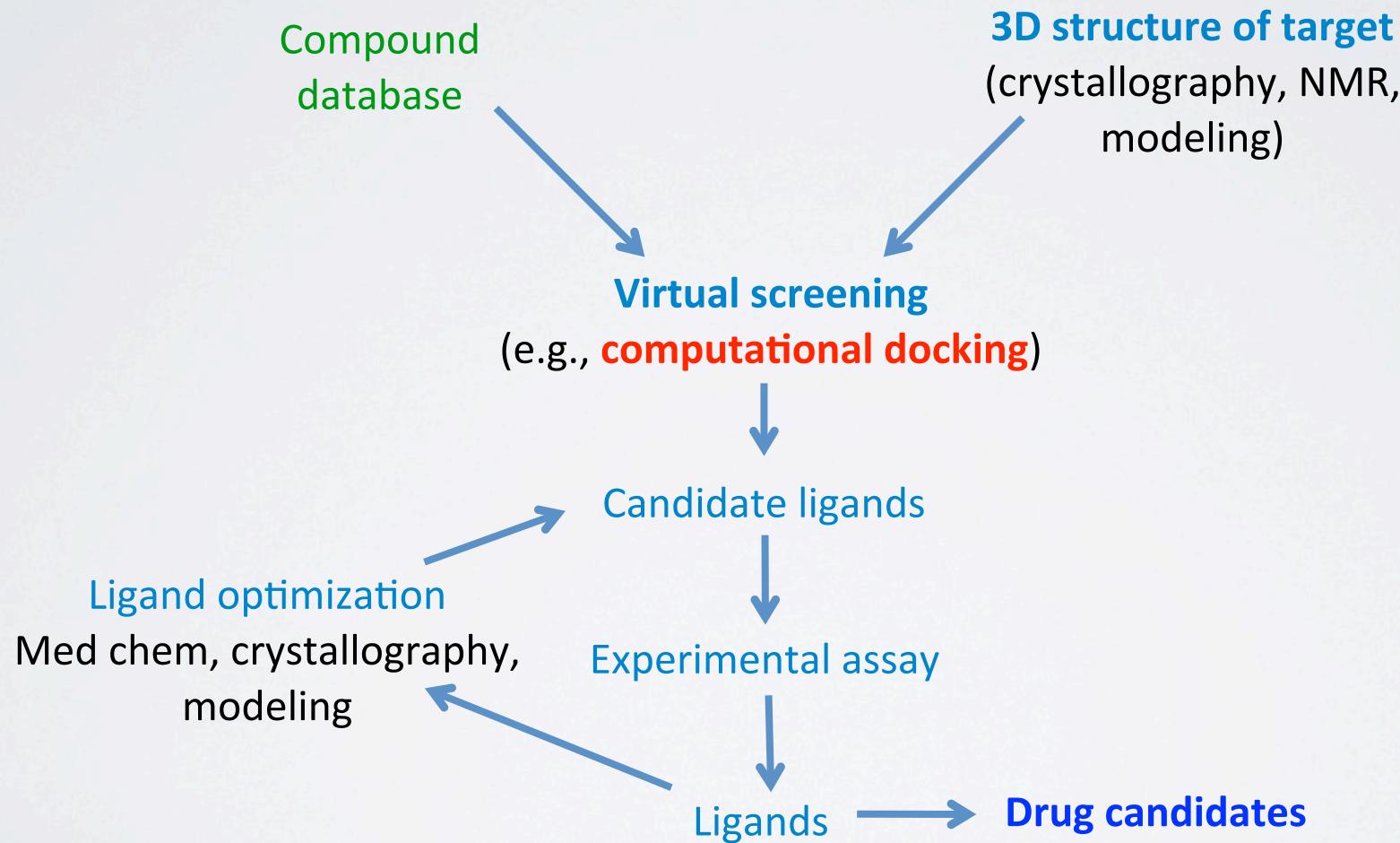


## Screened Coulombic



## Dihedral

# STRUCTURE-BASED VIRTUAL SCREENING



# COMPOUND LIBRARIES

The screenshot shows the Maybridge website. At the top, there's a navigation bar with links for Home, Building Blocks, Screening Libraries, Order, Tech Support, and About us. Below this is a banner for 'Thermo Fisher Scientific'. The main content area features a large image of a coastal landscape. On the left, there's a sidebar with a 'Search center' section containing icons for Building Blocks, Screening, and Structure search. The main content area has sections for 'Maybridge HitFinder™' (described as a pre-selected diverse screening library), 'Maximize quality hits from your screens', and 'Ready to Screen' (showing a 384-well plate). There's also a note about the HitFinder™ collection being pre-optimized for drug-likeness.

This image displays two websites side-by-side. On the left is the NIH Molecular Libraries Small Molecule Repository (MLSMR) website, which is part of the NIH Roadmap Initiative. It features a green header with the text 'NIH MOLECULAR LIBRARIES SMALL MOLECULE REPOSITORY' and 'A Roadmap Initiative'. The main content includes a 'Welcome' section with a photo of a scientist in a lab, information about the MLSMR project, and news items. On the right is the BioFocus website, which is a Galapagos company. It has a blue header with the BioFocus logo and 'A Galapagos Company'. The main content includes a 'Welcome' section with a photo of a scientist, information about BioFocus, and news items. Both sites have a footer with copyright information and links to other parts of the site.

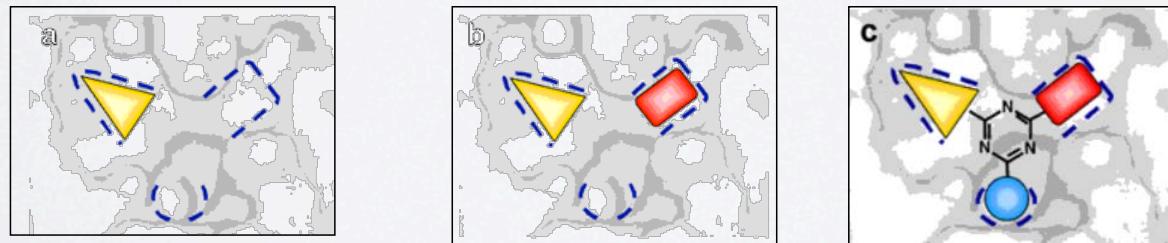
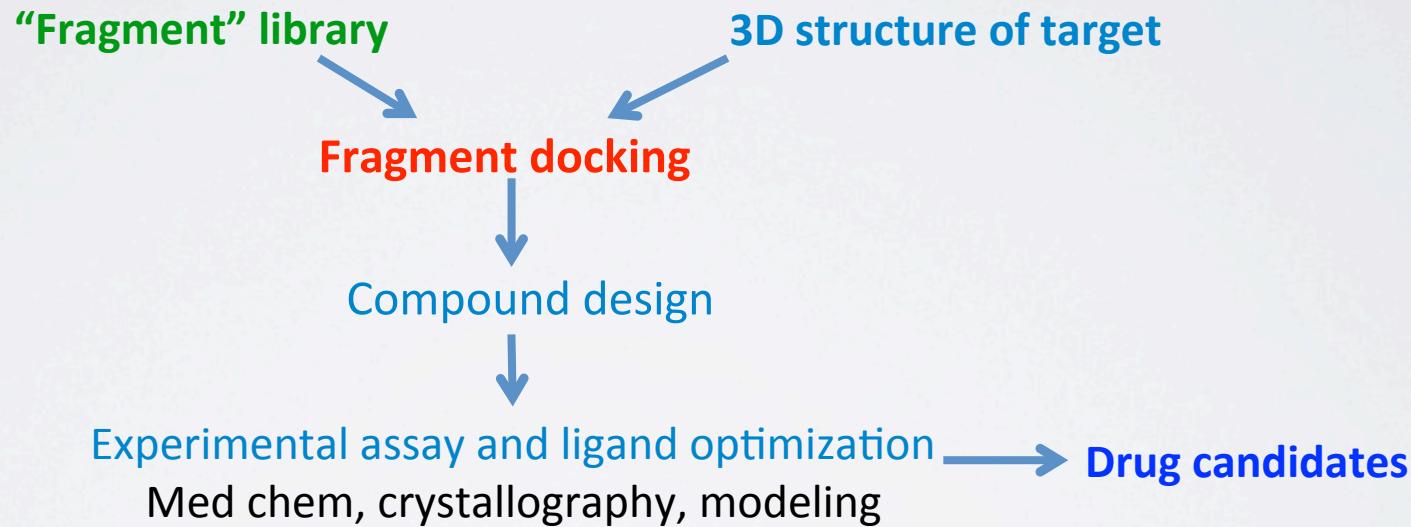
The Pittsburgh Molecular Libraries Screening Center (PMLSC) website is shown. The header includes the University of Pittsburgh logo, Pitt Home, Find People, and Contact Us. The main content features a large banner with the text 'PMLSC', 'BIG DISCOVERIES', and 'SMALL MOLECULES'. Below the banner is a 'Welcome' section with a photo of a scientist. The left sidebar contains links to various sections like Home, History, Personnel, Screening Technology, Compound Libraries, and more. The right sidebar has links for publications, contacts, and a keyword search. The footer includes links to Health Sciences @ Pitt, UPMC, NSSL, School of Medicine, Health Sciences Calendar, Our News & Events, and a note about the Office of the Senior Vice Chancellor for Health Sciences.

Commercial  
(in-house pharma)

Government (NIH)

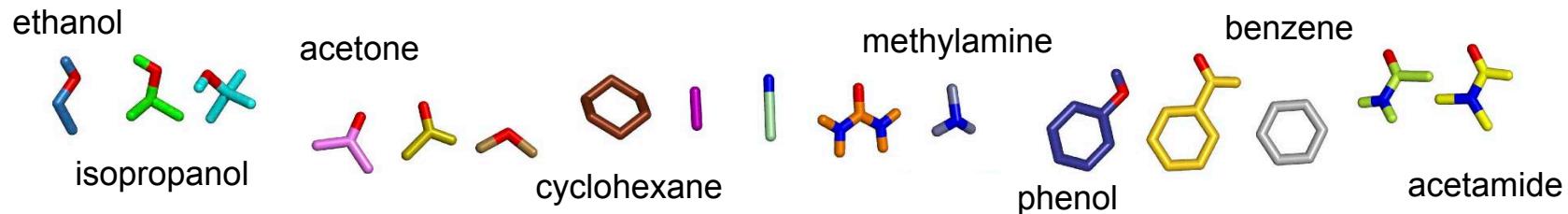
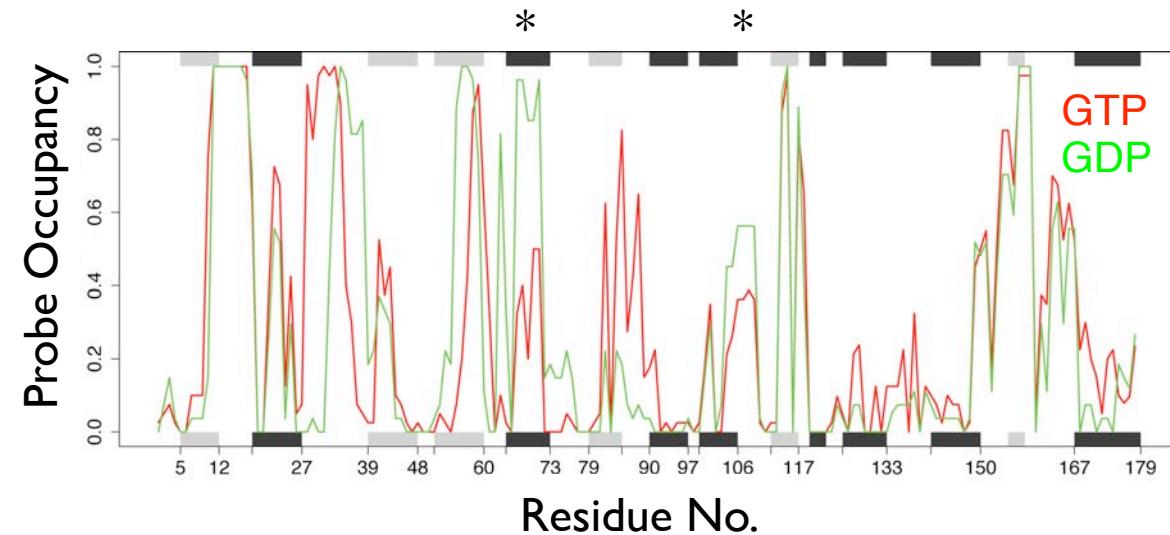
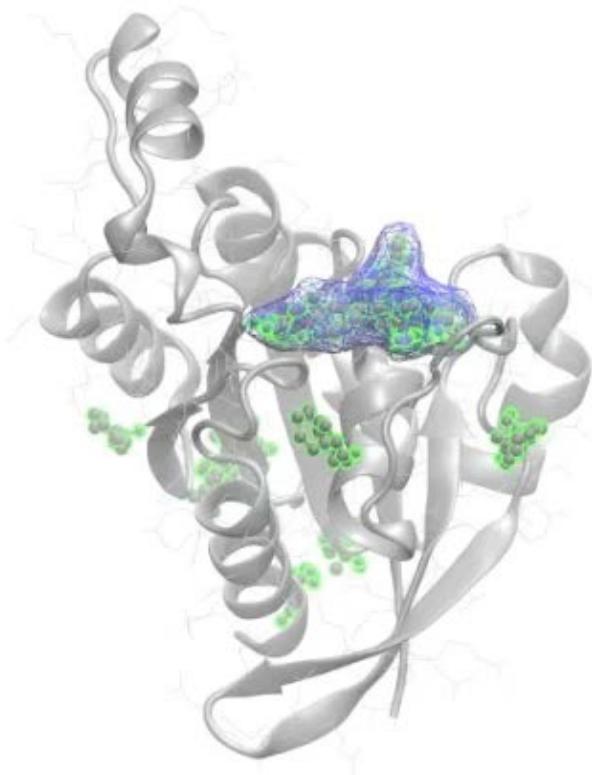
Academia

# FRAGMENTAL STRUCTURE-BASED SCREENING



# Multiple non active-site pockets identified

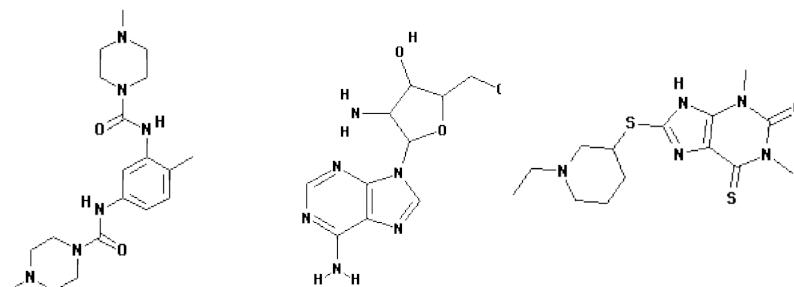
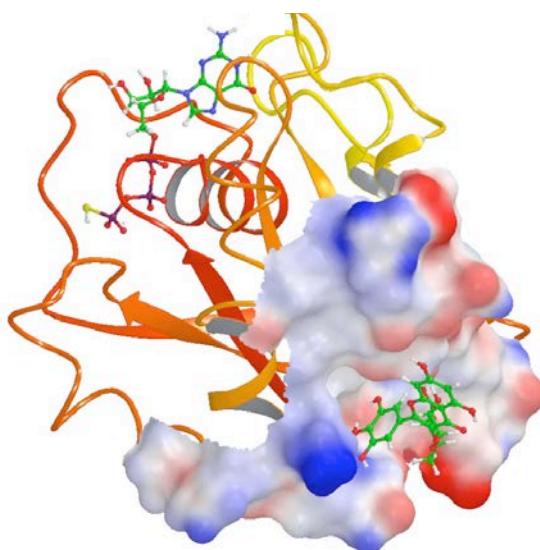
Small organic probe fragment affinities map multiple potential binding sites across the structural ensemble.



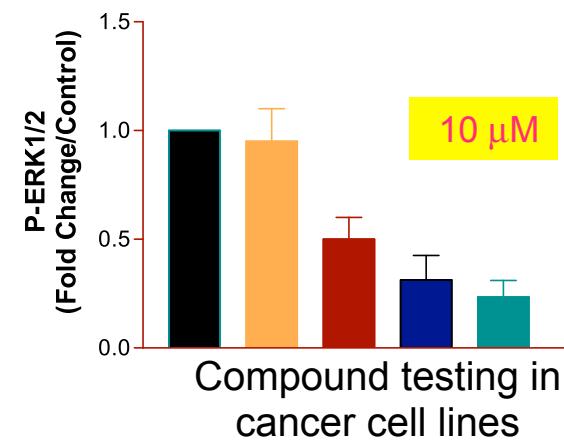
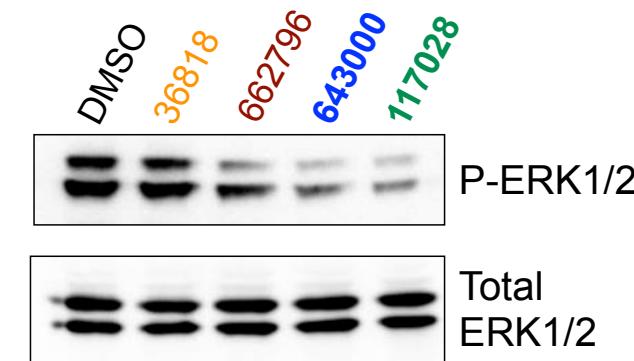
# Ensemble docking & candidate inhibitor testing

Top hits from ensemble docking against distal pockets were tested for inhibitory effects on basal ERK activity in glioblastoma cell lines.

Ensemble computational docking

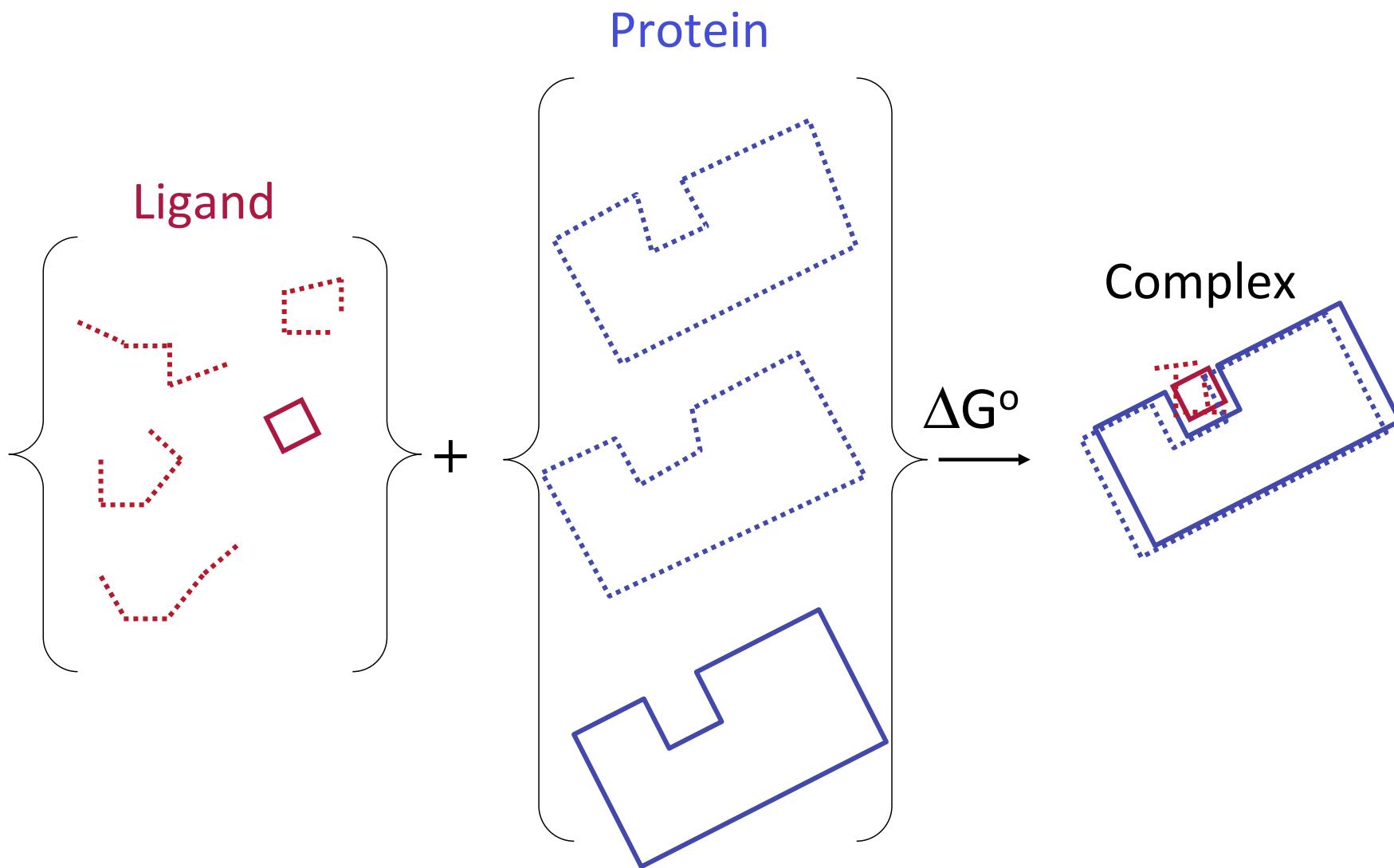


Compound effect on U251 cell line



PLoS One (2011, 2012)

# Proteins and Ligand are Flexible



# COMMON SIMPLIFICATIONS USED IN PHYSICS-BASED DOCKING

Quantum effects approximated classically

Protein often held rigid

Configurational entropy neglected

Influence of water treated crudely

Two main approaches:

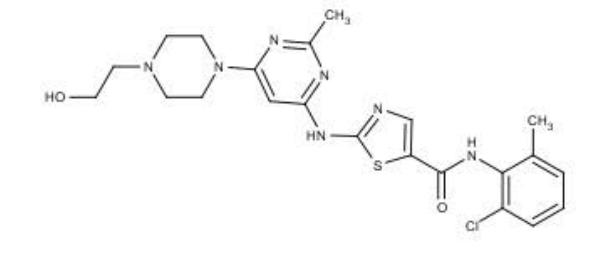
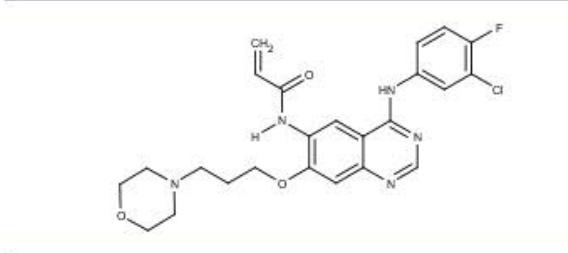
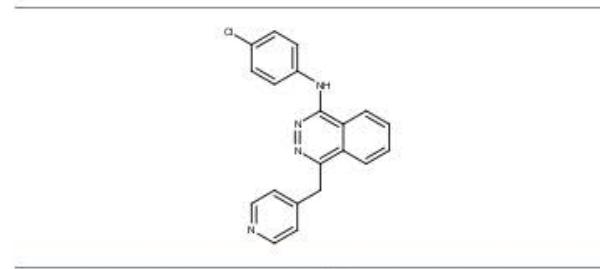
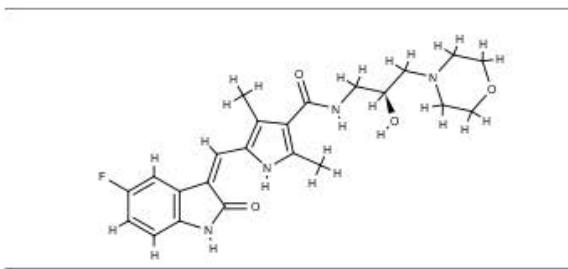
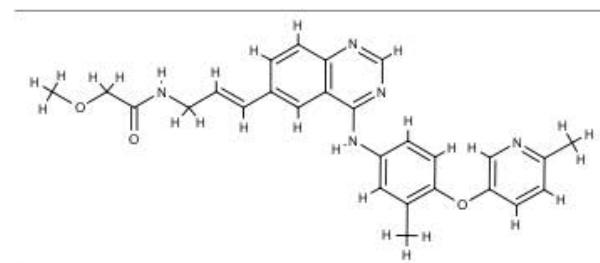
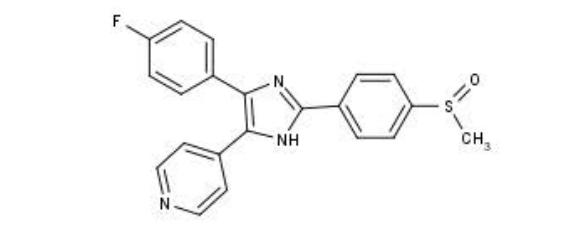
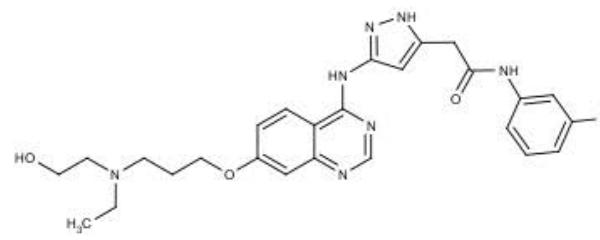
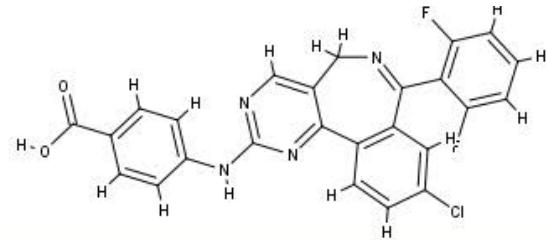
(1). Receptor/Target-Based

(2). Ligand/Drug-Based

# Scenario 2

Structure of Targeted Protein Unknown: Ligand-Based Drug Discovery

e.g. MAP Kinase Inhibitors



Using knowledge of existing inhibitors to discover more

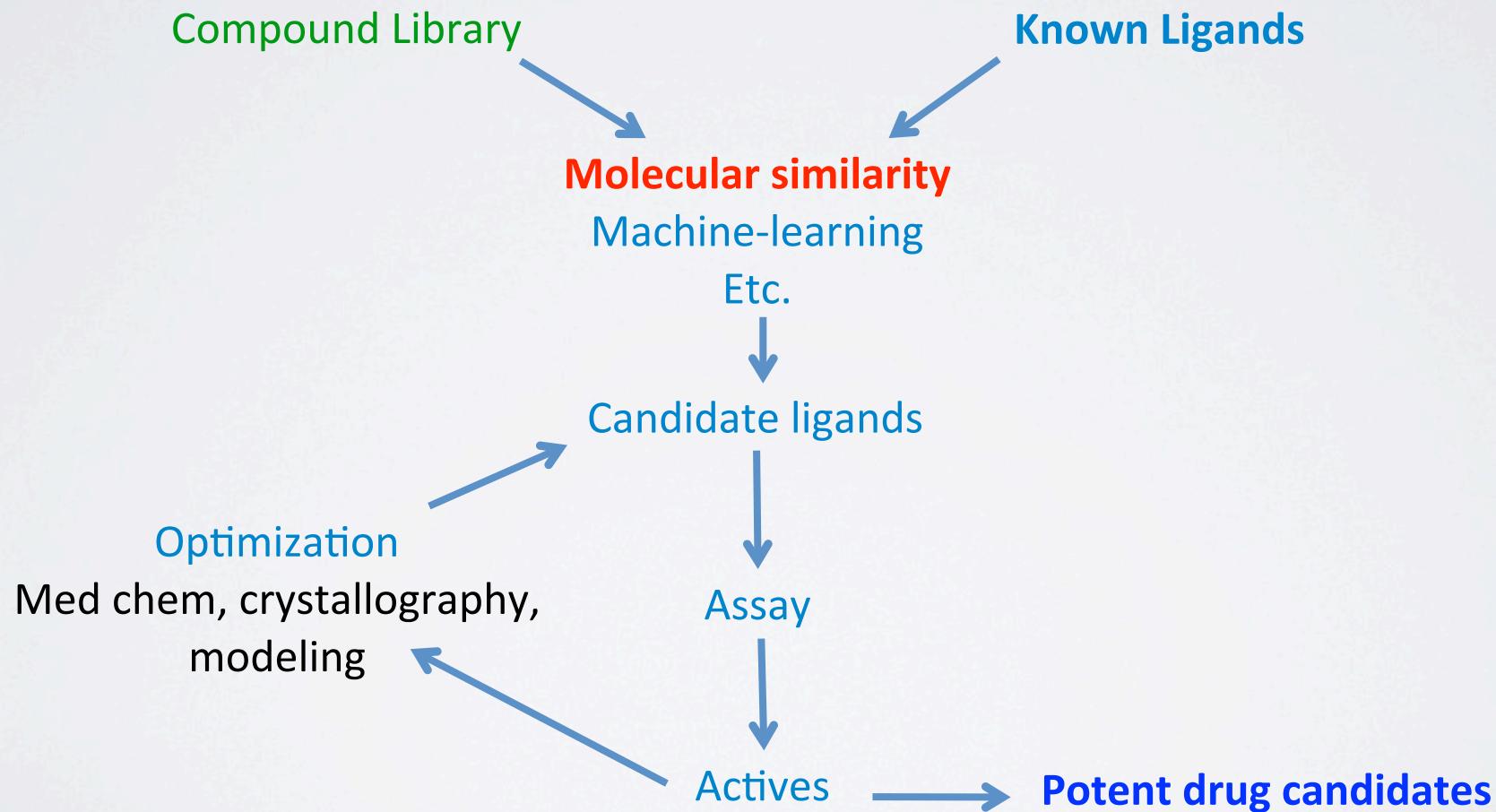
## Why Look for Another Ligand if You Already Have Some?

Experimental screening generated some ligands, but they don't bind tightly

A company wants to work around another company's chemical patents

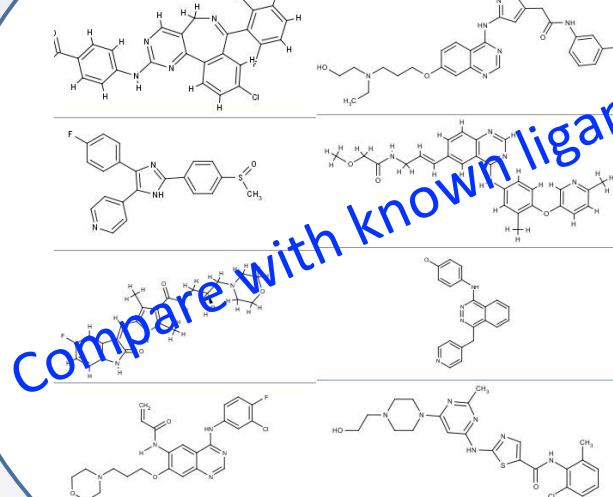
An high-affinity ligand is toxic, is not well-absorbed, etc.

# LIGAND-BASED VIRTUAL SCREENING



# CHEMICAL SIMILARITY LIGAND-BASED DRUG-DISCOVERY

Compounds  
(available/synthesizable)



Different

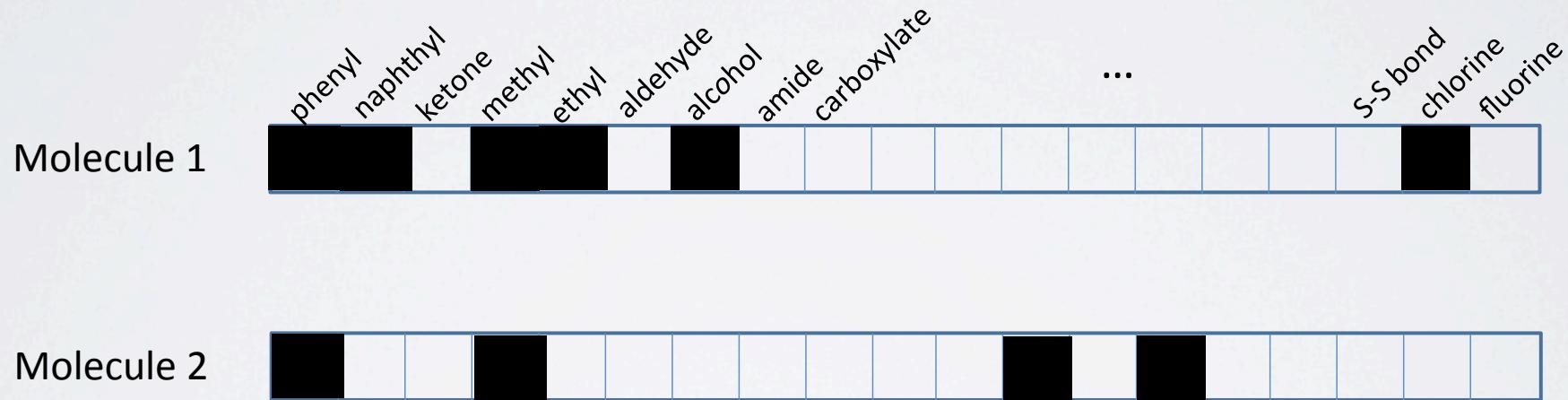
Don't bother

Similar

Test experimentally

# CHEMICAL FINGERPRINTS

## BINARY STRUCTURE KEYS



# CHEMICAL SIMILARITY FROM FINGERPRINTS

Tanimoto Similarity  
or Jaccard Index, T

$$T \equiv \frac{N_I}{N_U} = 0.25$$

Intersection



$N_I=2$

Union



$N_U=8$

Molecule 1



Molecule 2



# POTENTIAL DRAWBACKS OF PLAIN CHEMICAL SIMILARITY

May miss good ligands by being overly conservative

Too much weight on irrelevant details

# Abstraction and Identification of Relevant Compound Features

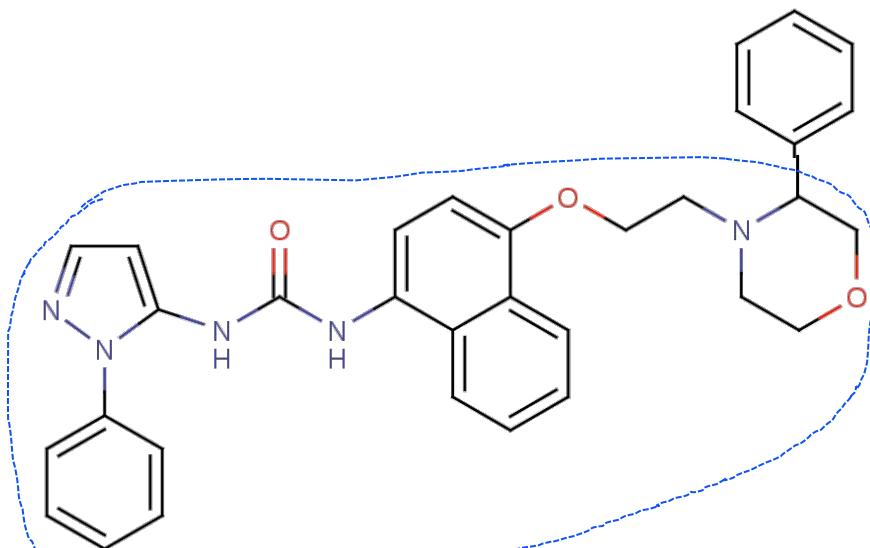
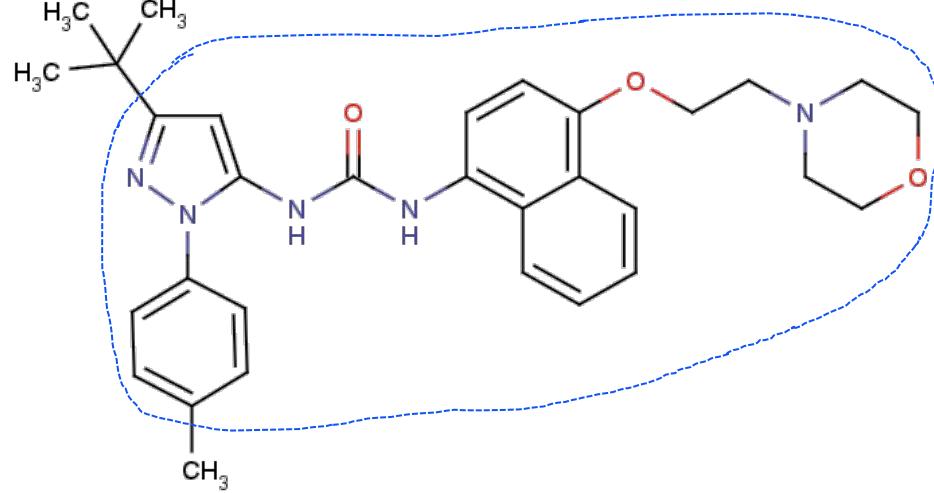
Ligand shape and common substructures

Pharmacophore models

Chemical descriptors

Statistics and machine learning

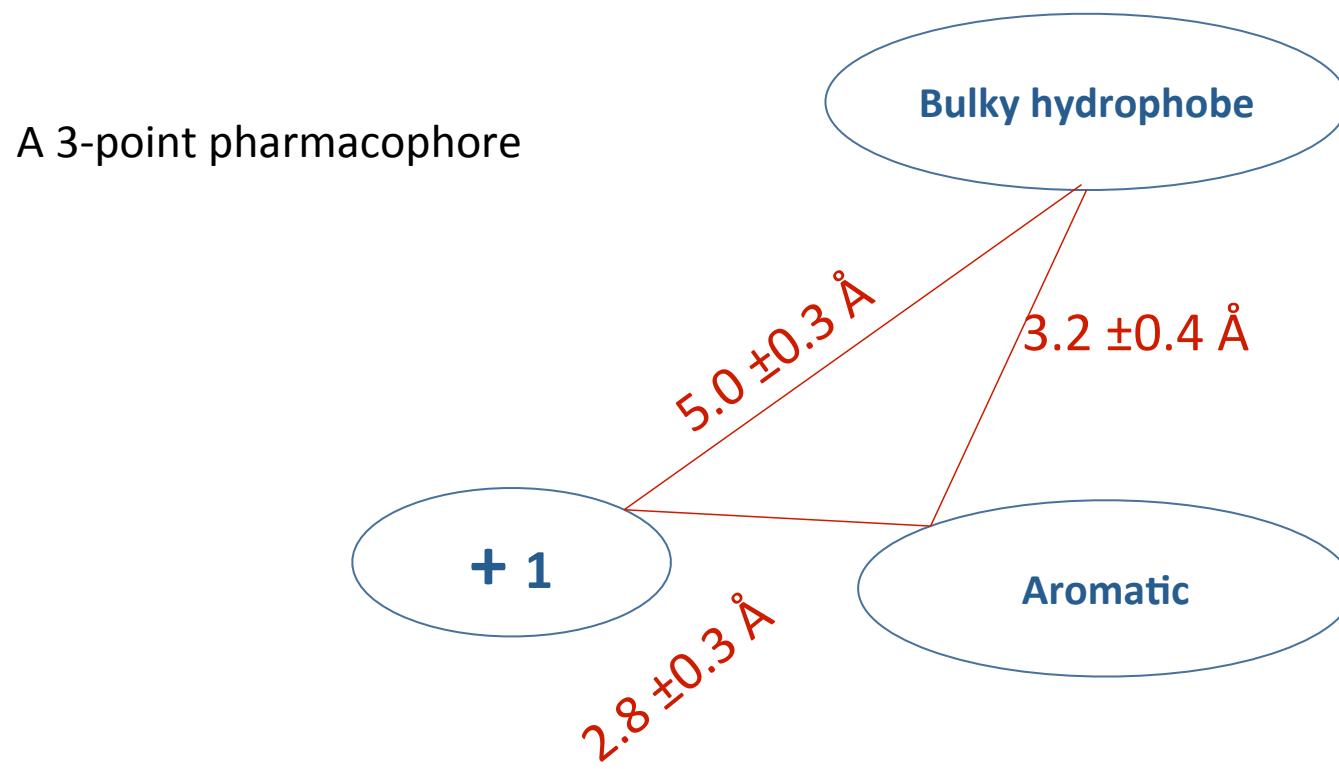
# Maximum Common Substructure



$N_{\text{common}} = 34$

# Pharmacophore Models

Φάρμακο (drug) + Φορά (carry)



# Molecular Descriptors

## More abstract than chemical fingerprints

### Physical descriptors

molecular weight

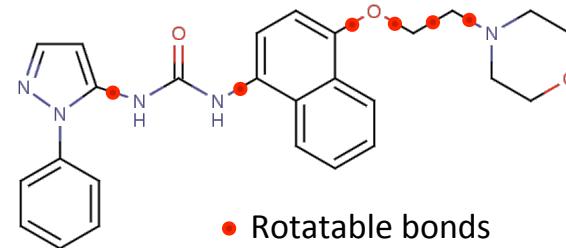
charge

dipole moment

number of H-bond donors/acceptors

number of rotatable bonds

hydrophobicity ( $\log P$  and  $c\log P$ )



### Topological

branching index

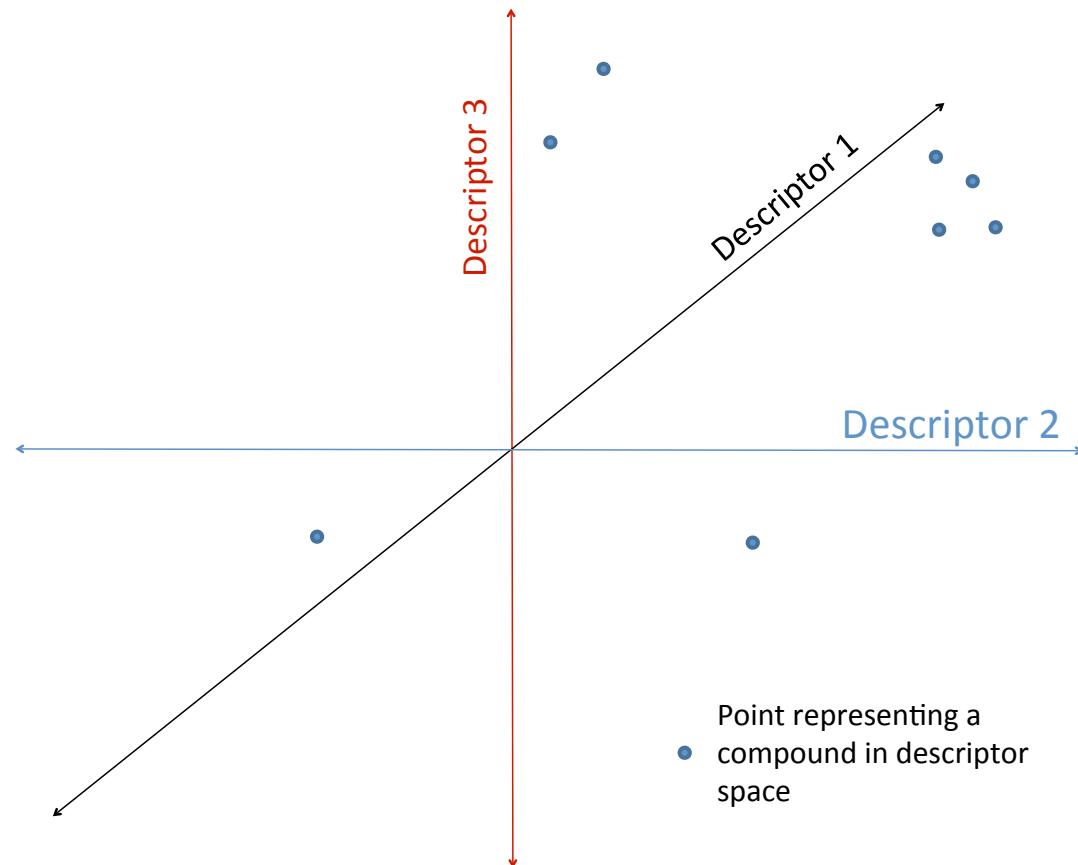
measures of linearity vs interconnectedness

Etc. etc.

# A High-Dimensional “Chemical Space”

Each compound is at a point in an n-dimensional space

Compounds with similar properties are near each other



# Statistics and Machine Learning

## Some examples

Partial least squares

Support vector machines

Genetic algorithms for descriptor-selection

# Summary

Overview of drug discovery

Computer-aided methods

Structure-based

Ligand-based

Interaction potentials

Physics-based

Knowledge-based (data driven)

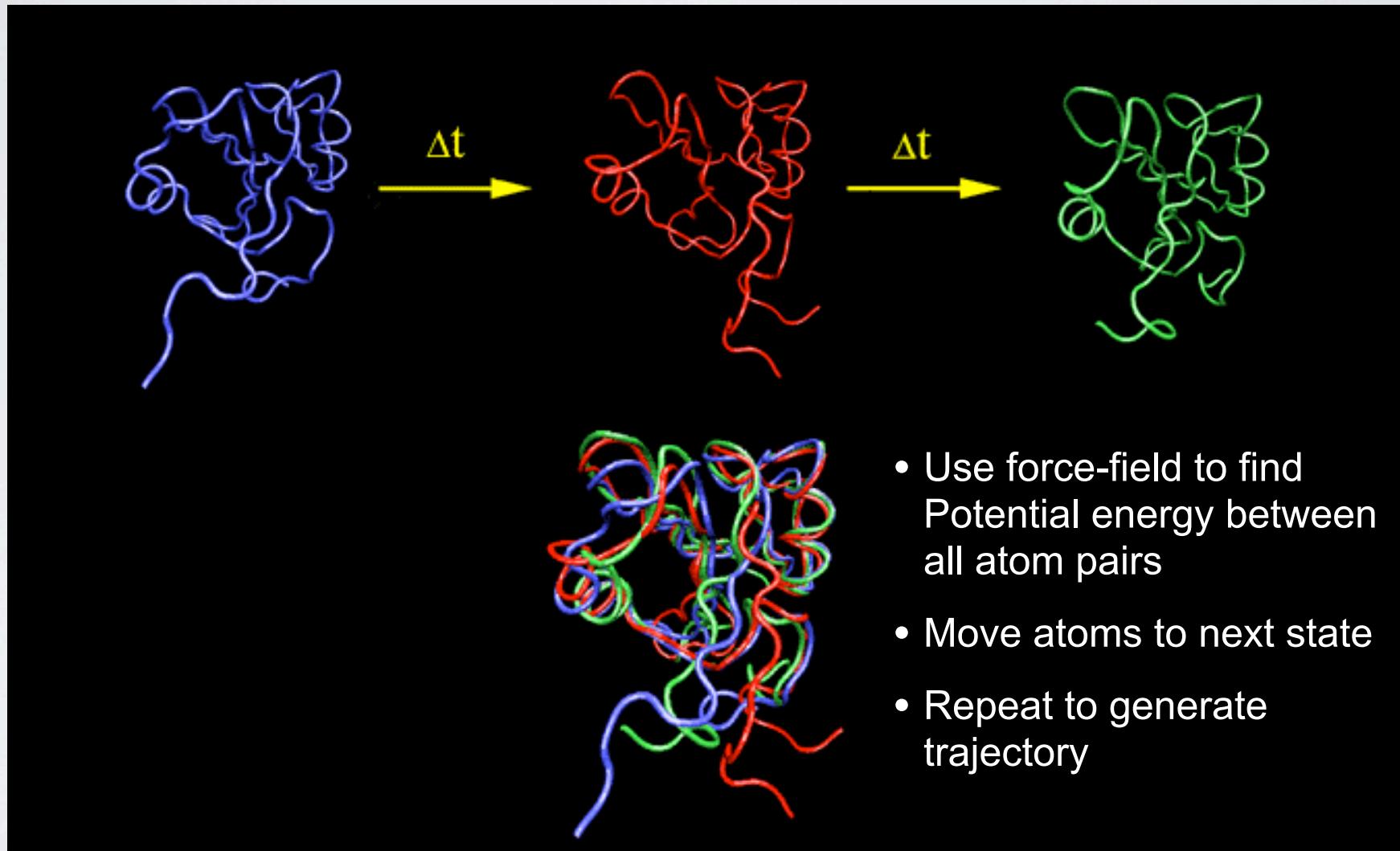
Ligand-protein databases, machine-readable chemical formats

Ligand similarity and beyond

# PREDICTING FUNCTIONAL DYNAMICS MOLECULAR DYNAMICS SIMULATIONS

- Proteins are intrinsically flexible molecules with internal motions that are often intimately coupled to their biochemical function.
  - E.g. ligand and substrate binding, allosteric regulation
- Thus knowledge of dynamics can provide a deeper understanding of the mapping of structure to function.
- Molecular dynamics (MD) and normal mode analysis (NMA) are two major methods for predicting and characterizing molecular motions

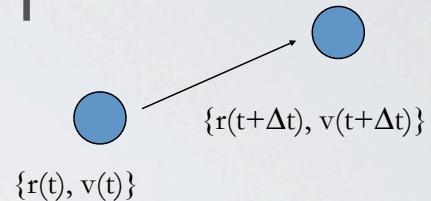
# Molecular Dynamics Simulation



McCammon, Gelin & Karplus, *Nature* (1977)

# MD ALGORITHM

- Initialize system
  - (Randomly) assign velocities.
  - Find the potential energy between all atom pairs
- Move and integrate equations of motion.
  - Find new velocities and positions
- Repeat

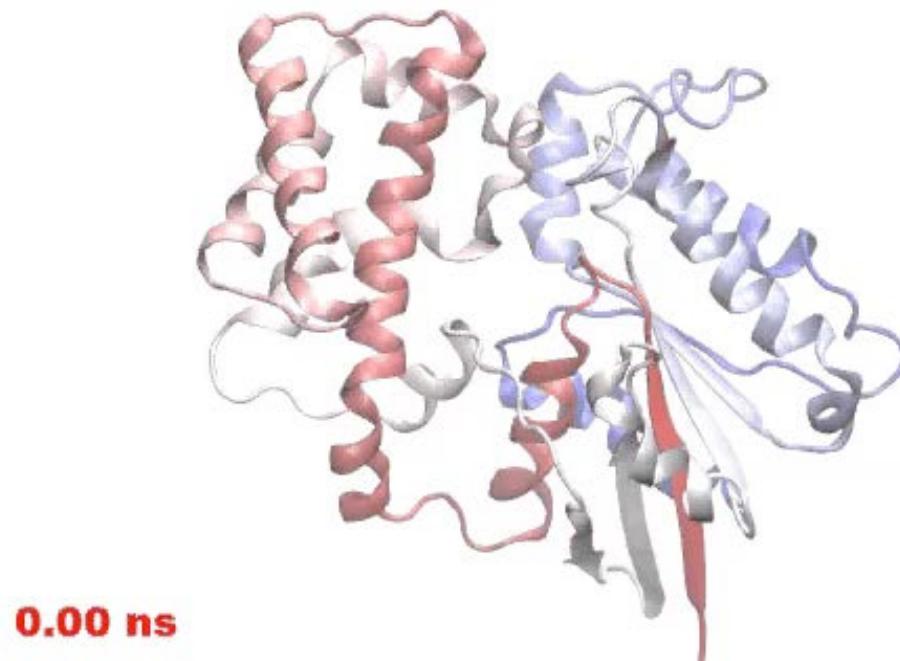


*Leapfrog algorithm*

- |  |  |
|--|--|
| <p>→ 1 solve for <math>a_i</math> at <math>t</math> using:</p> <p>2 update <math>v_i</math> at <math>t + \Delta t/2</math> using:</p> <p>3 update <math>r_i</math> at <math>t + \Delta t</math> using:</p> | $-\frac{dE}{dr_i} = F_i = m_i \mathbf{a}_i(t)$ $v_i(t + \Delta t/2) = v_i(t - \Delta t/2) + a_i(t) \Delta t$ $r_i(t + \Delta t) = r_i(t) + v_i(t + \Delta t/2) \Delta t$ |
|--|--|

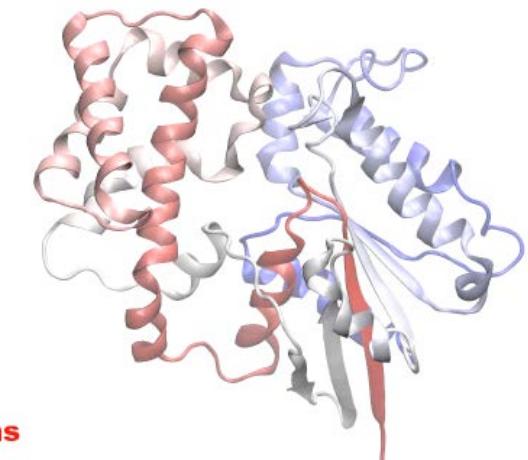
# MD Prediction of Functional Motions

Accelerated MD simulation of  
nucleotide-free transducin alpha subunit

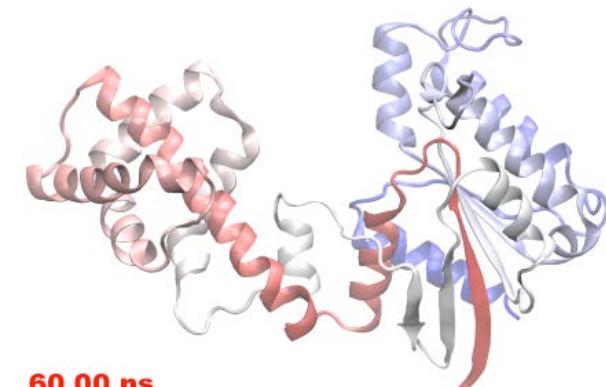


Yao and Grant, Biophys J. (2013)

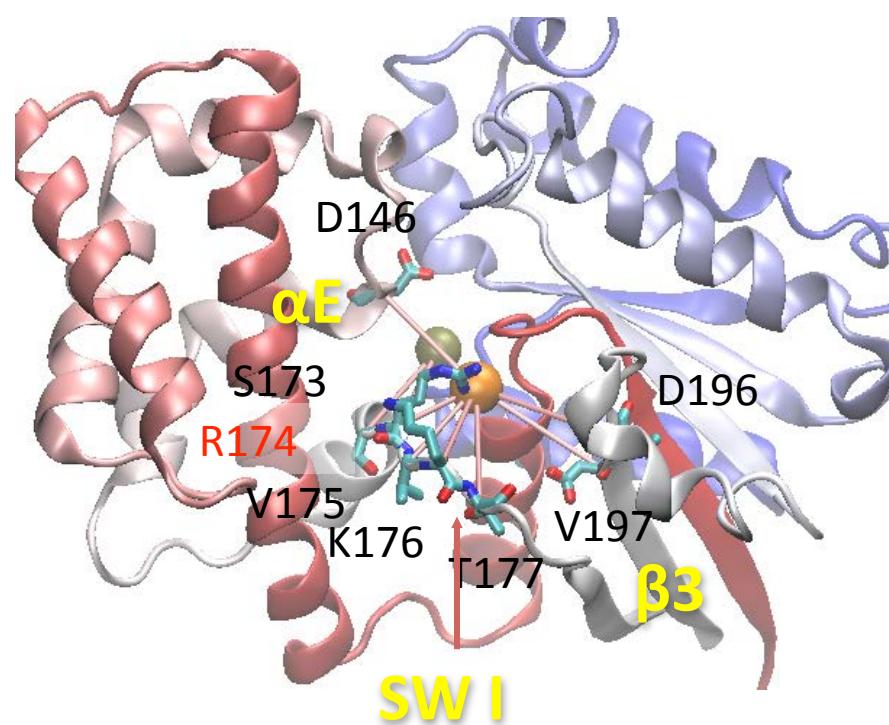
“close”



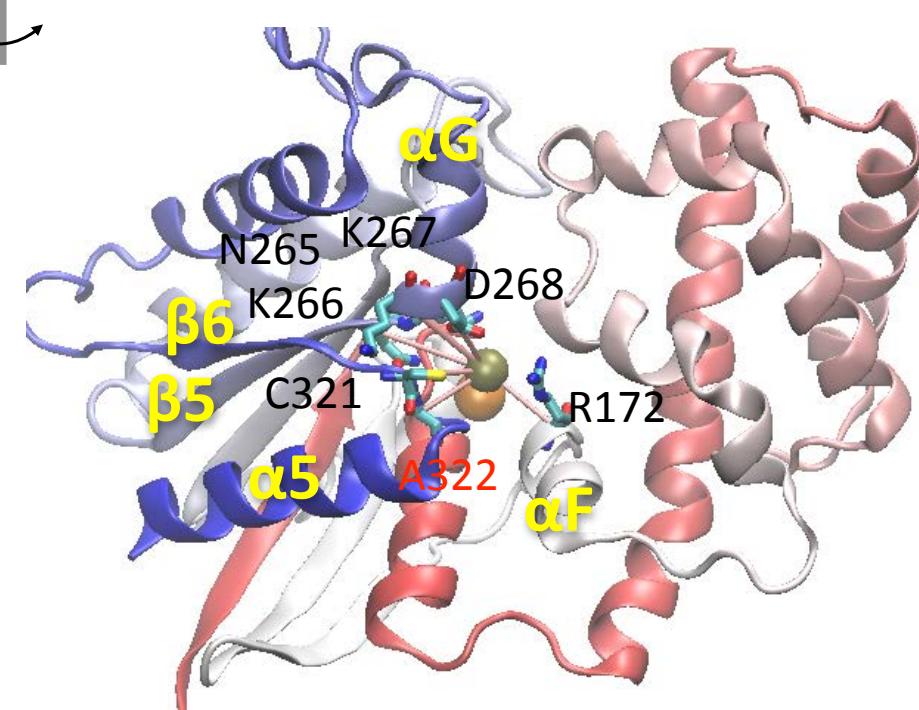
“open”



# Key Residues Mediating Coupling Between Residues And Nucleotide



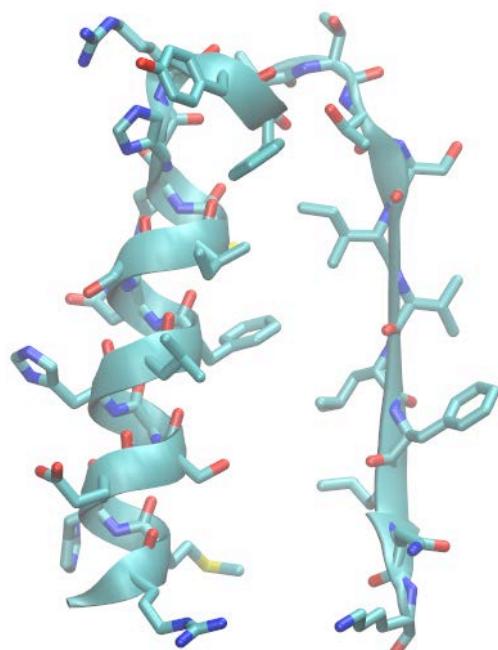
180°



Yao and Grant, Biophys J. (2013)

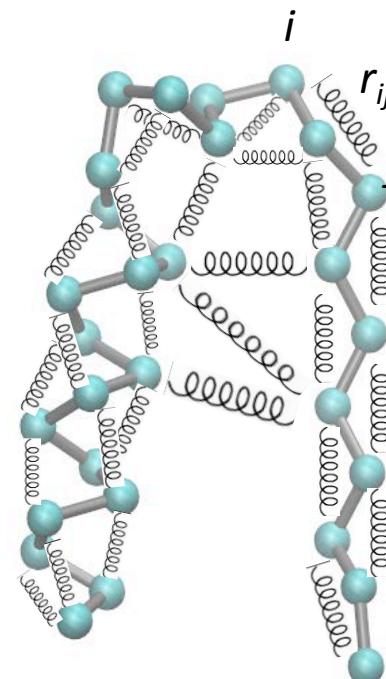
# Normal Mode Analysis (NMA)

- Accelerated MD is still time-consuming
- Elastic network model (ENM)
  - Finish in **seconds!**



Atomic

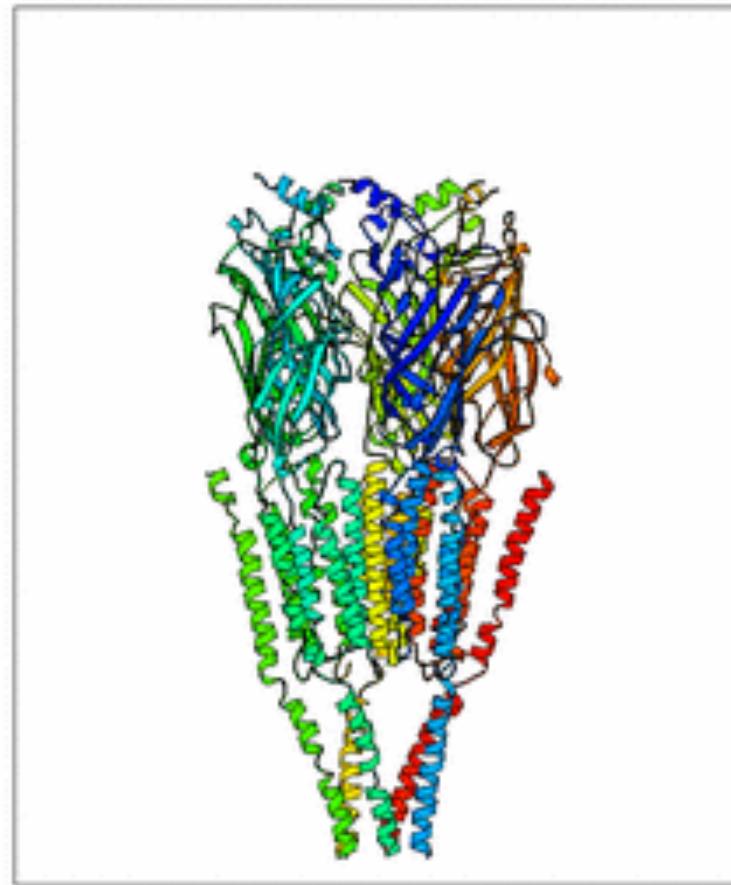
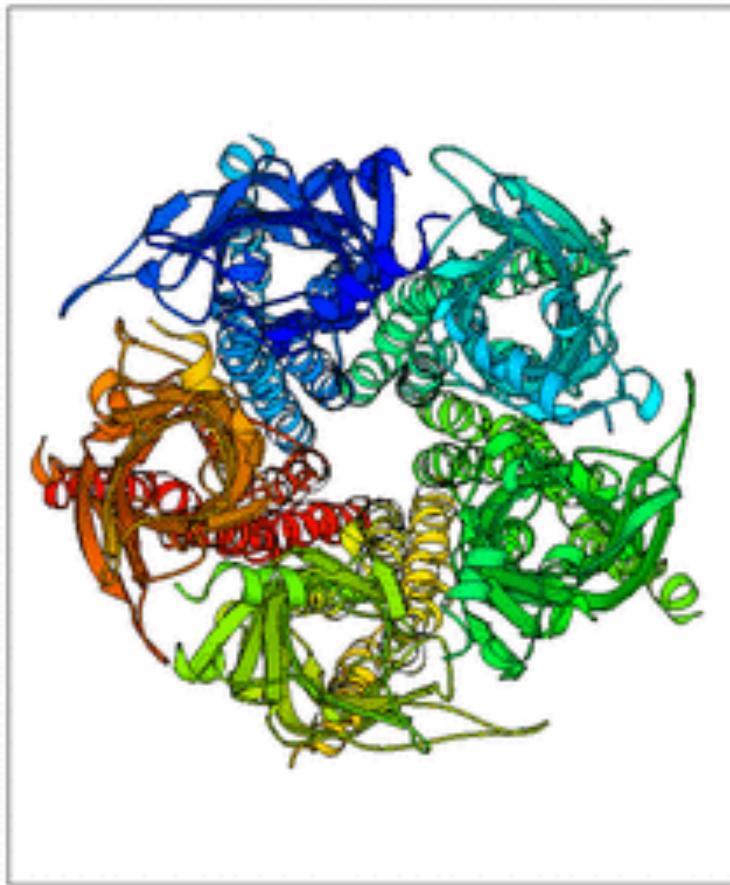
C. G.



a. a.

- 1 bead / 1 amino acid
- Connected by springs

# Normal mode of acetylcholine receptor



- The receptor displays an twist like motion, responsible for the axially symmetric opening and closing of the ion channel

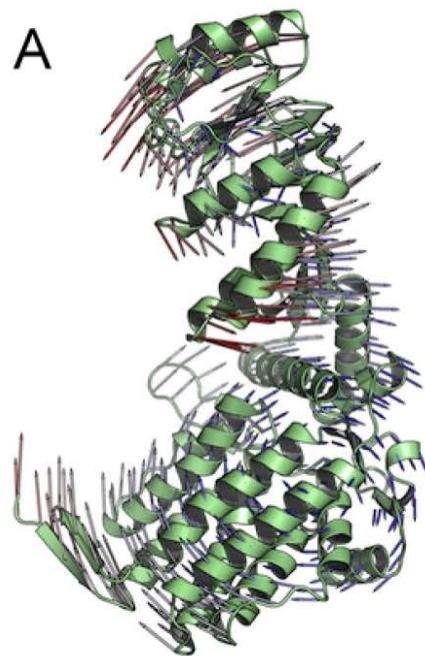
# Problems in Conventional ENM-NMA

## Conventional ENM-NMA

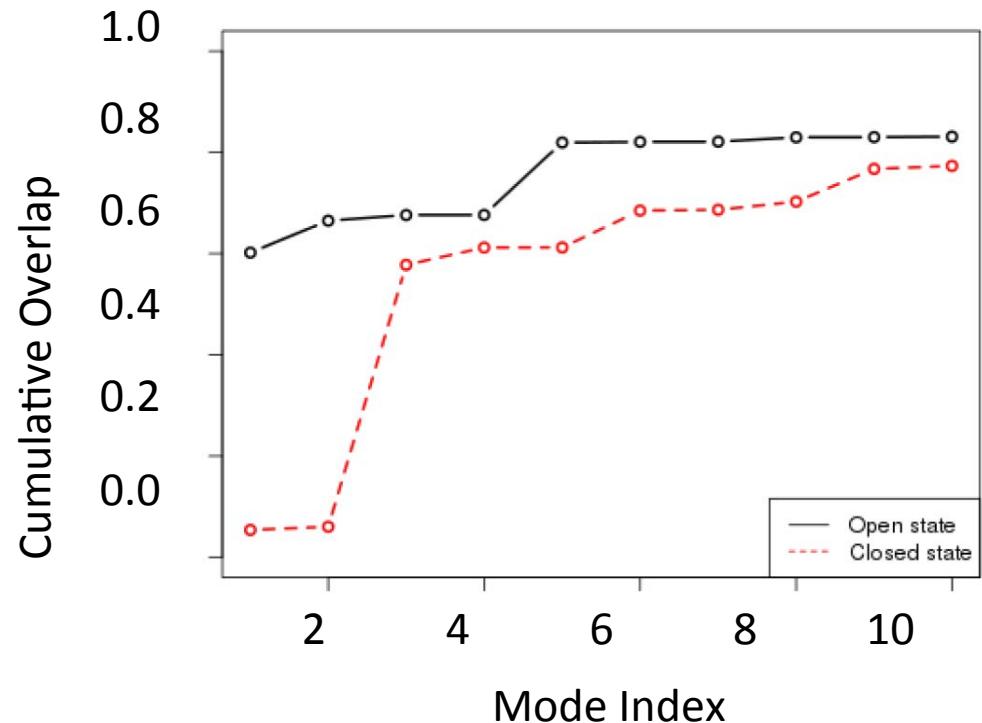
• Work well for elongated multi-domain systems such as GroEL

• But, results are **dependent** on the **input structure** - open forms work best!

GroEL

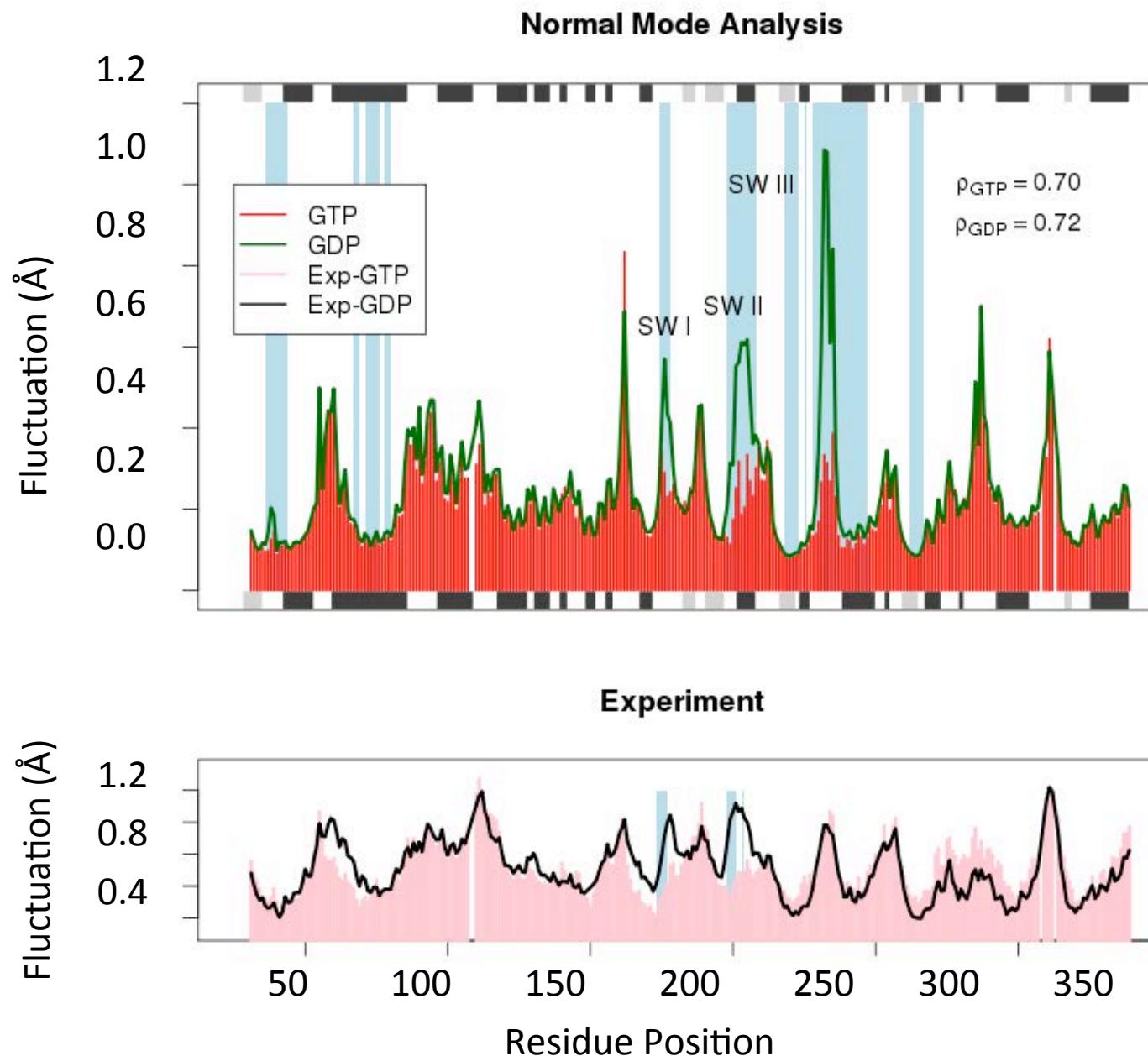


B



**Overlap:** Dot product of modes and position difference vector between open and close states

# NMA Predicts High Flexibility in Functional Regions



# SUMMARY

- Structural bioinformatics is computer aided structural biology
- Structural data plays a central role in bioinformatics
- Reviewed the fundamentals of protein structure
- Introduced both physics and knowledge based modeling approaches for describing the structure, energetics and dynamics of proteins computationally
- Described common applications in drug design and for prediction of functional motions.

# INFORMING SYSTEMS BIOLOGY?

