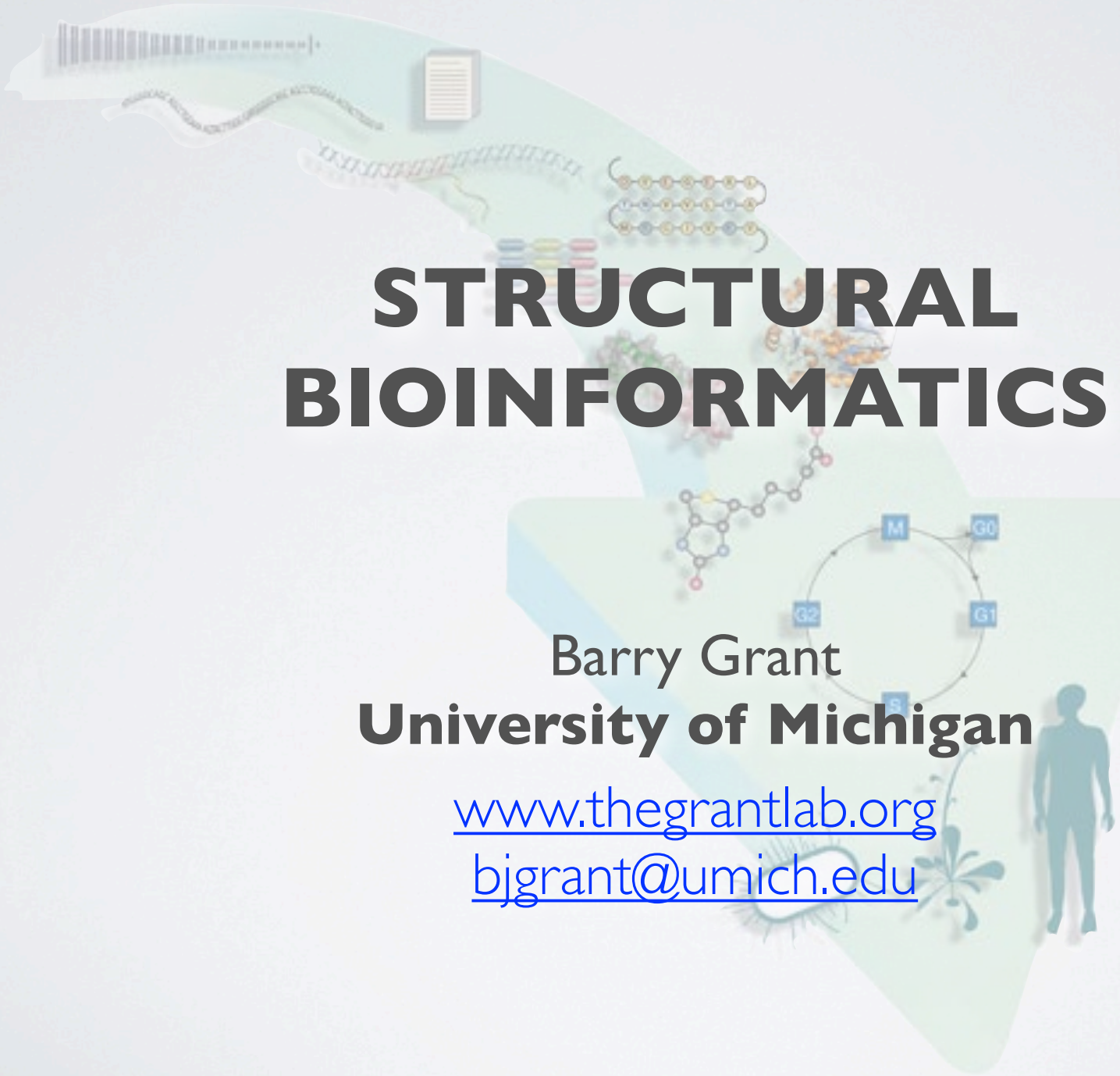


STRUCTURAL BIOINFORMATICS

Barry Grant
University of Michigan
www.thegrantlab.org
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?

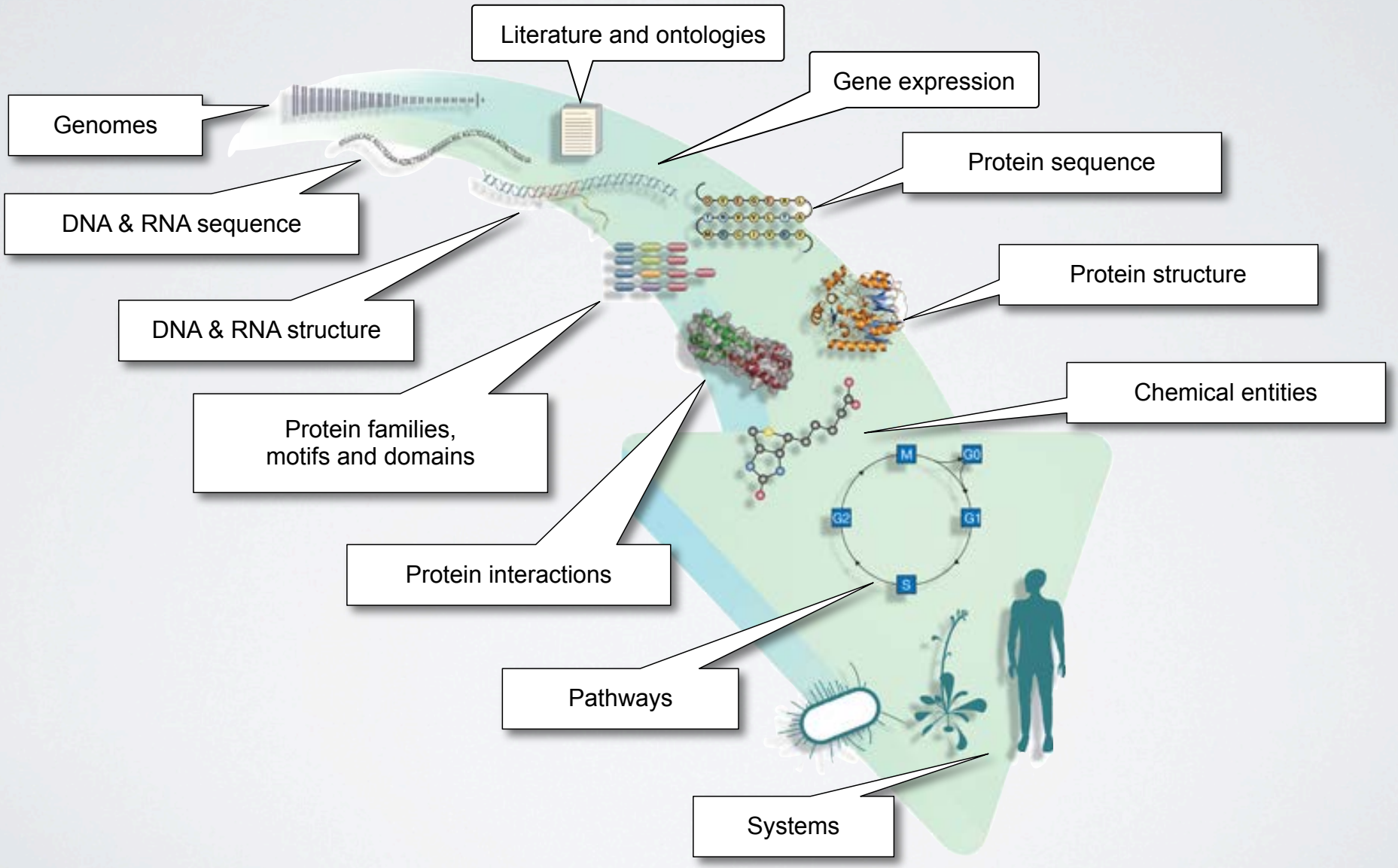
- **Structural bioinformatics is computer aided structural biology!**
- Characterizes biomolecules and their assembles at the molecular & atomic level.

Q. Why should we care?

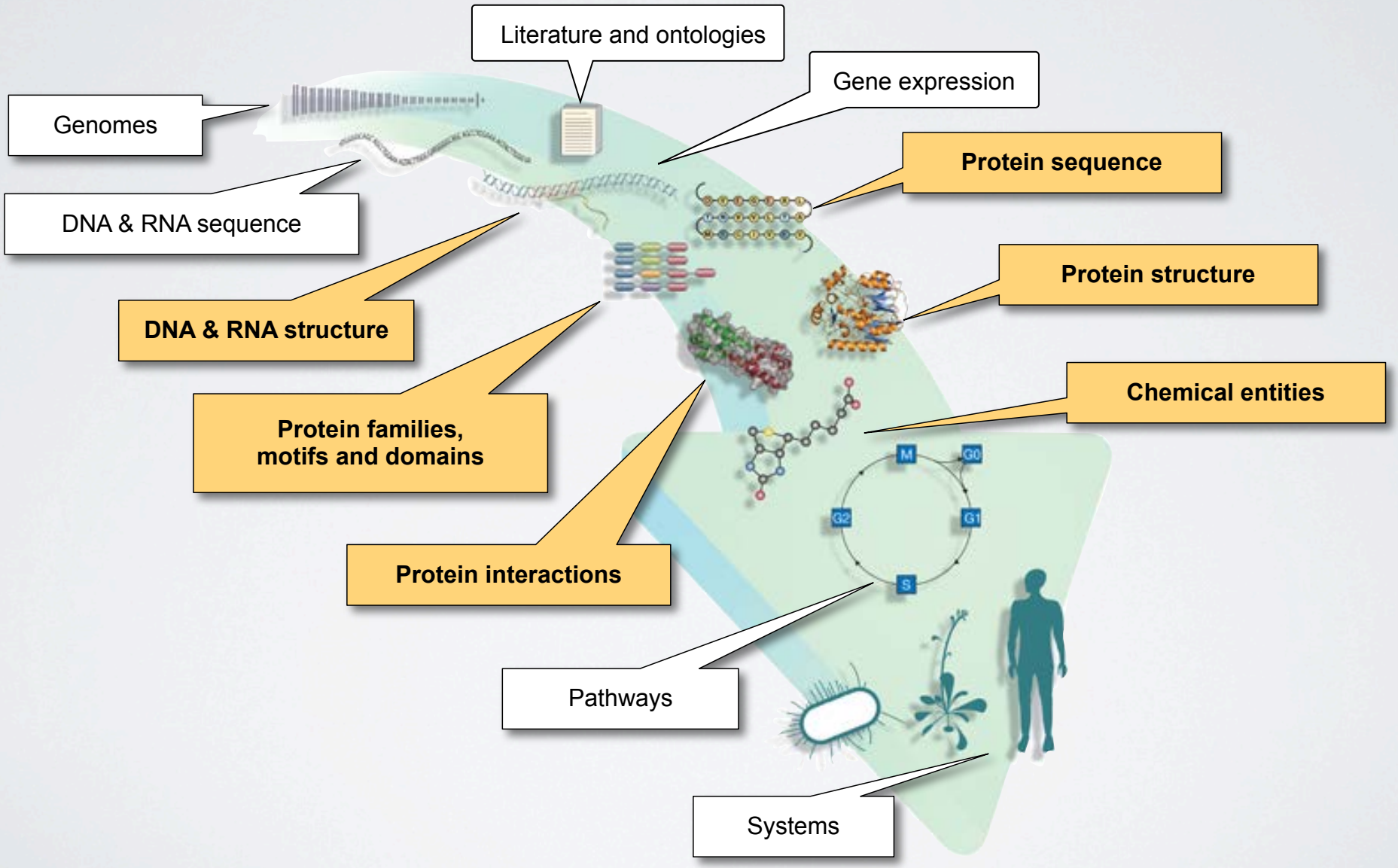
- Because biomolecules are "*nature's robots*" [Tanford, 2001]

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

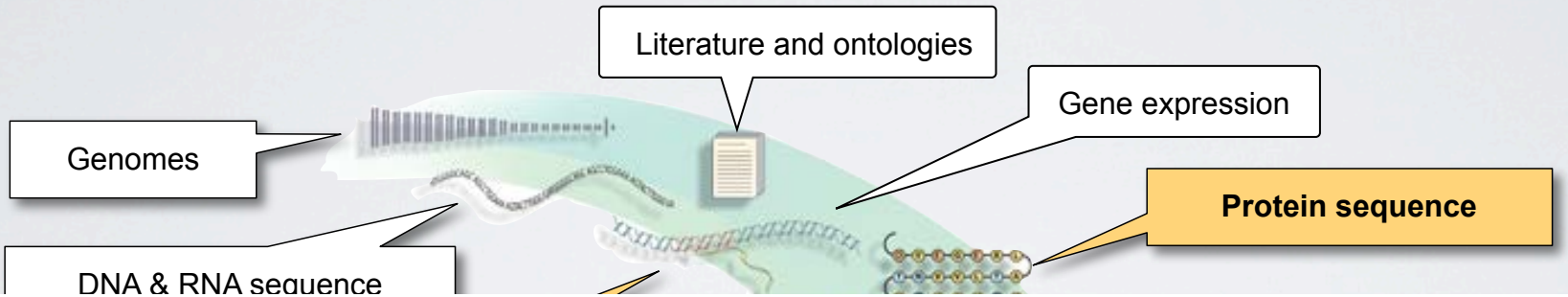
BIOINFORMATICS DATA



STRUCTURAL DATA IS CENTRAL

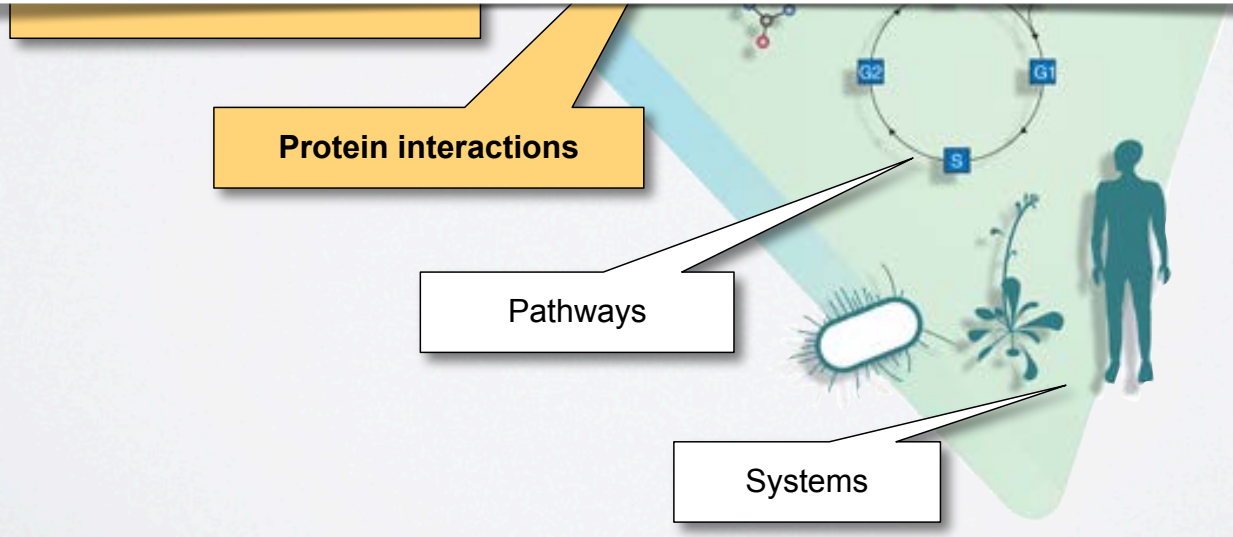


STRUCTURAL DATA IS CENTRAL

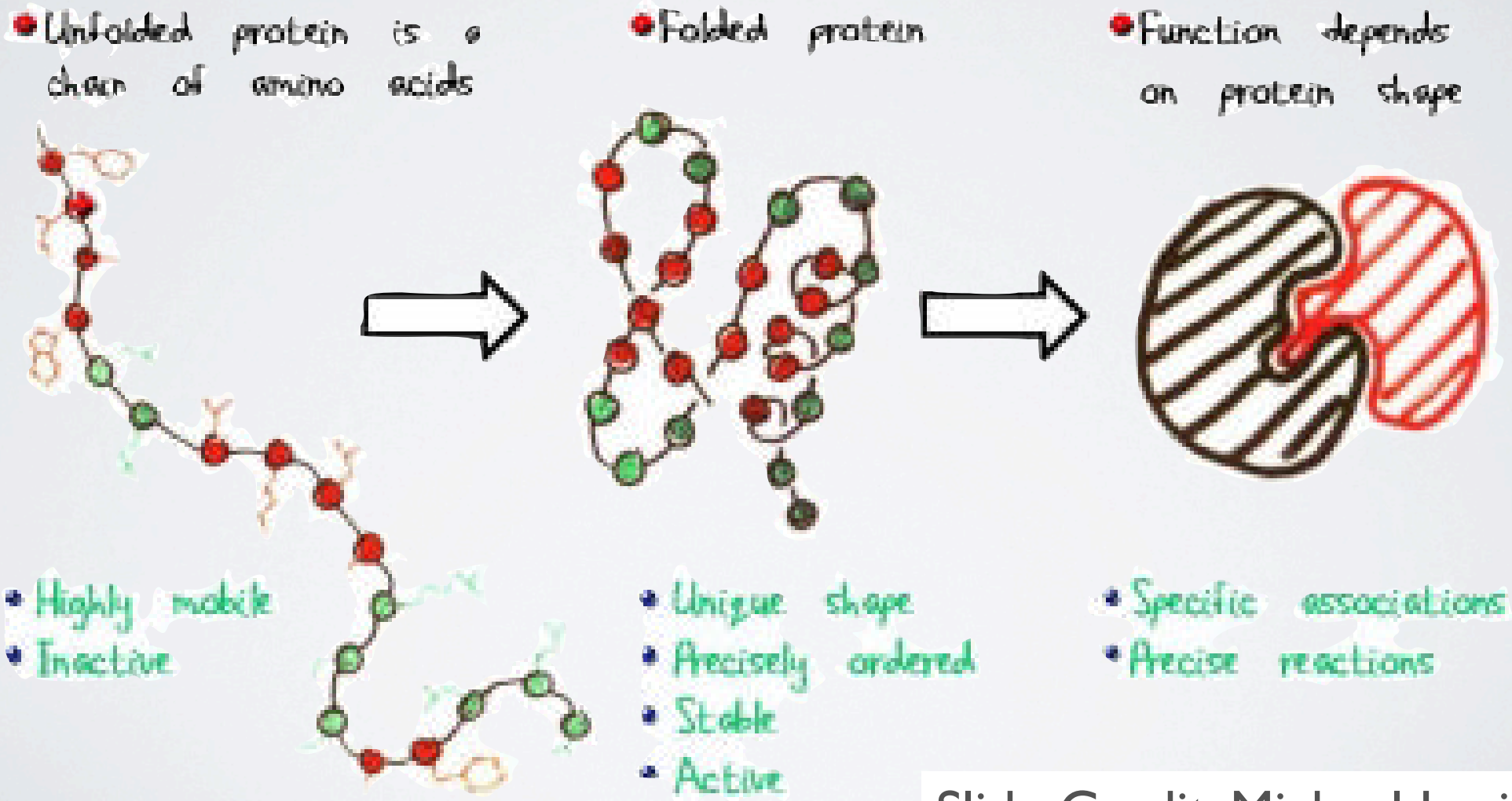


THE HOLY TRINITY OF STRUCTURAL BIOINFORMATICS

Sequence > Structure > Function

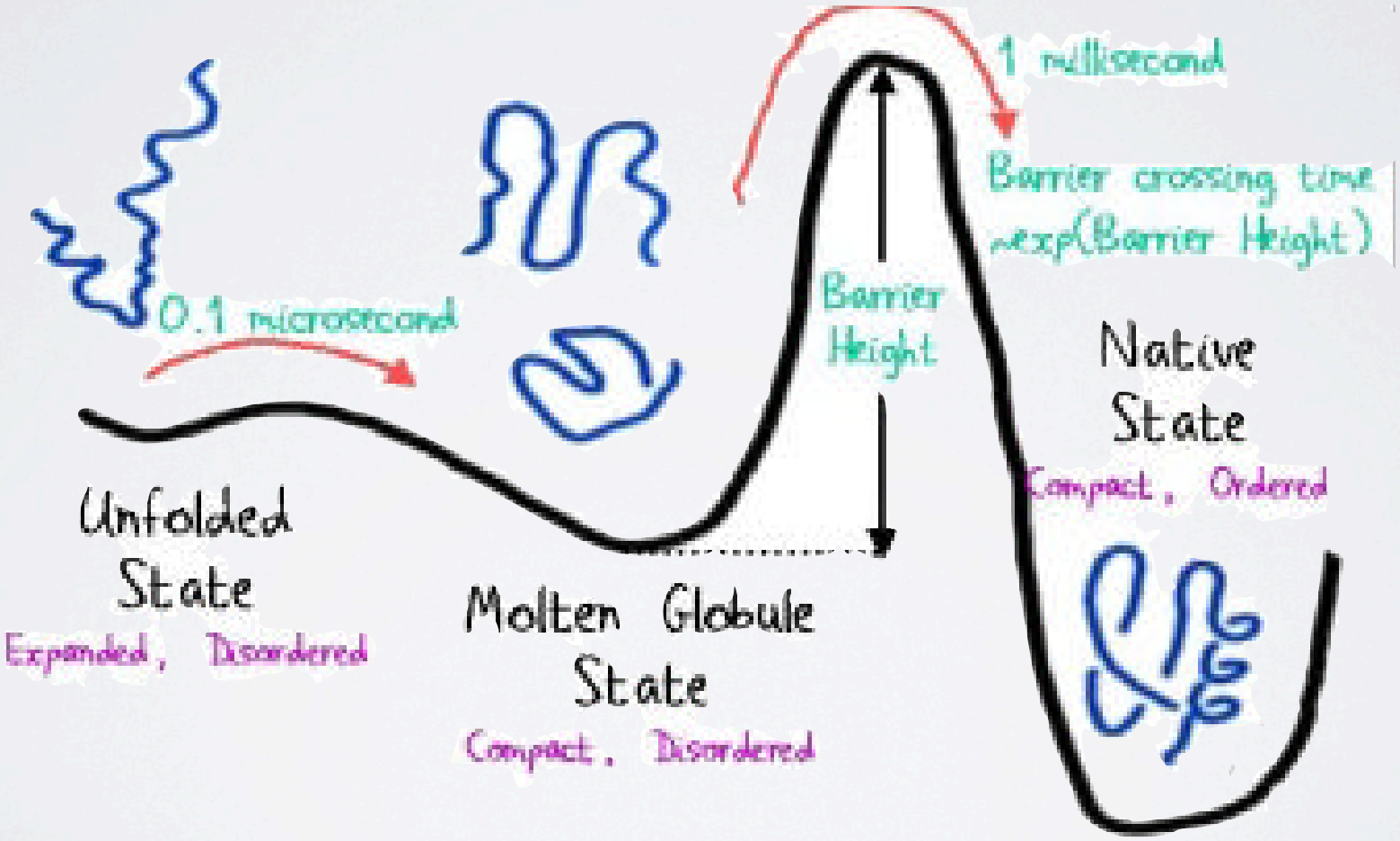


Sequence > Structure > Function



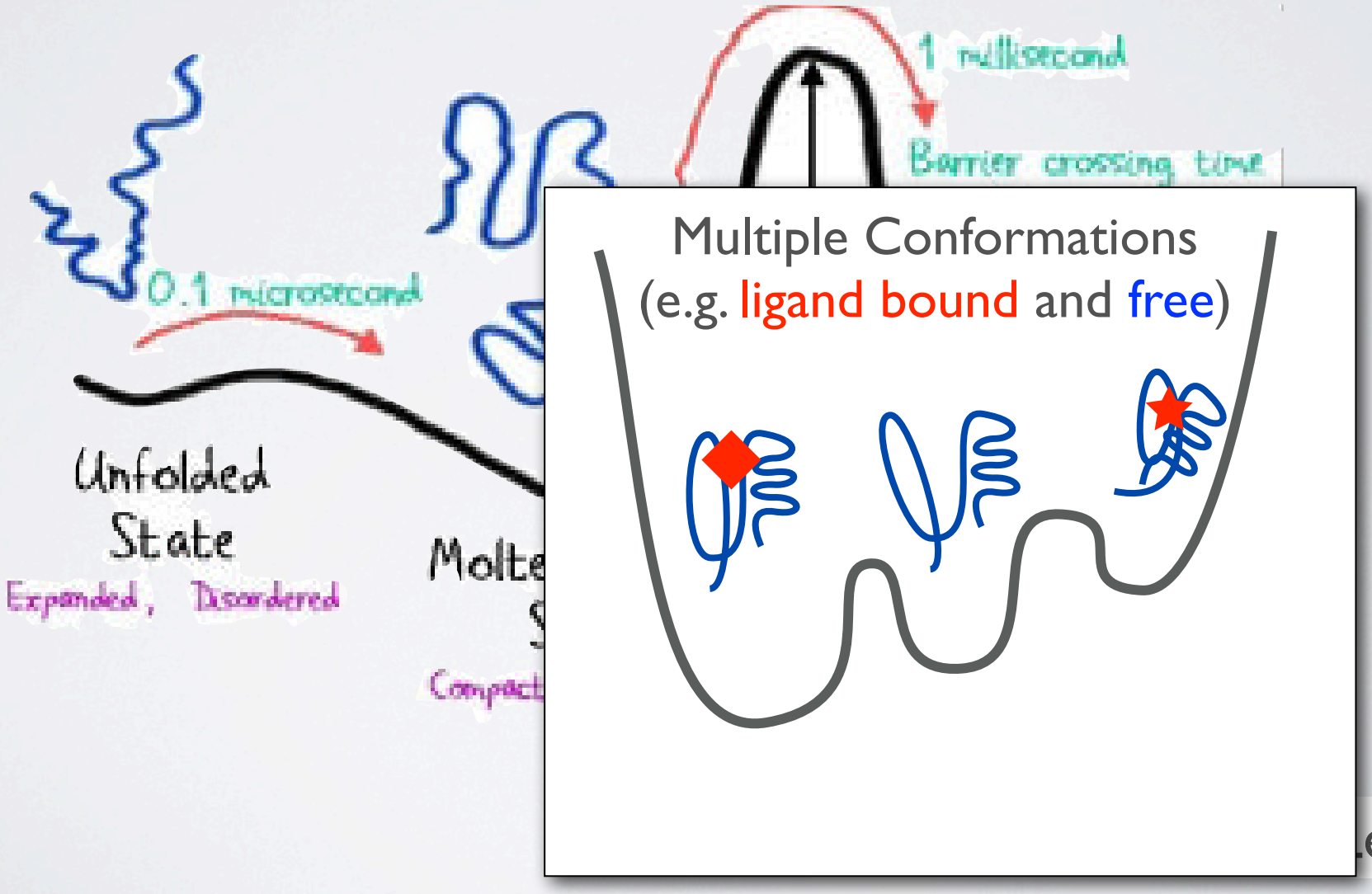
Slide Credit: Michael Levitt

KEY CONCEPT: **ENERGY LANDSCAPE**



Slide Credit: Michael Levitt

KEY CONCEPT: **ENERGY LANDSCAPE**



evitt

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
- **Representin**
  - of structure
  - 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

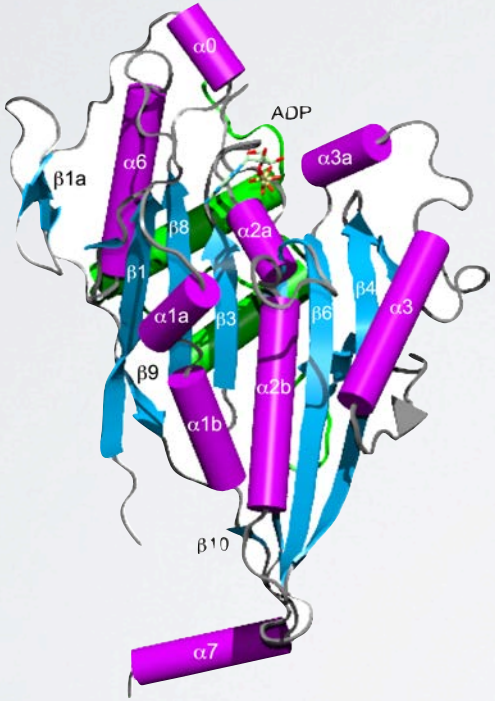
Next Lecture:

- Predicting structure from sequence [Prof. Zhang]

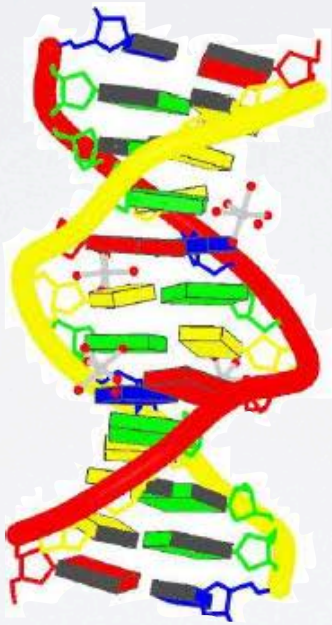
TODAY'S MENU:

- **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

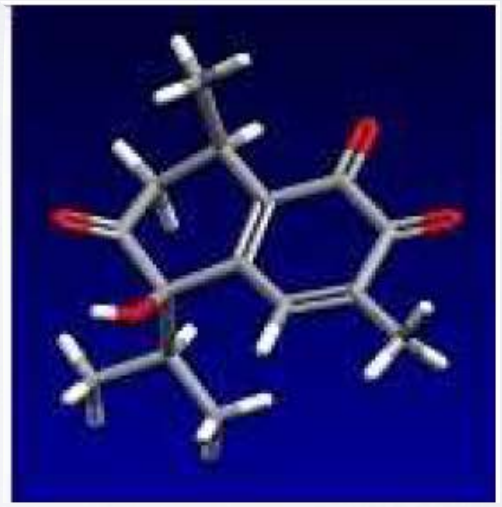
TRADITIONAL FOCUS **PROTEIN**, **DNA** AND **SMALL MOLECULE** DATA SETS WITH **MOLECULAR STRUCTURE**



Protein
(PDB)



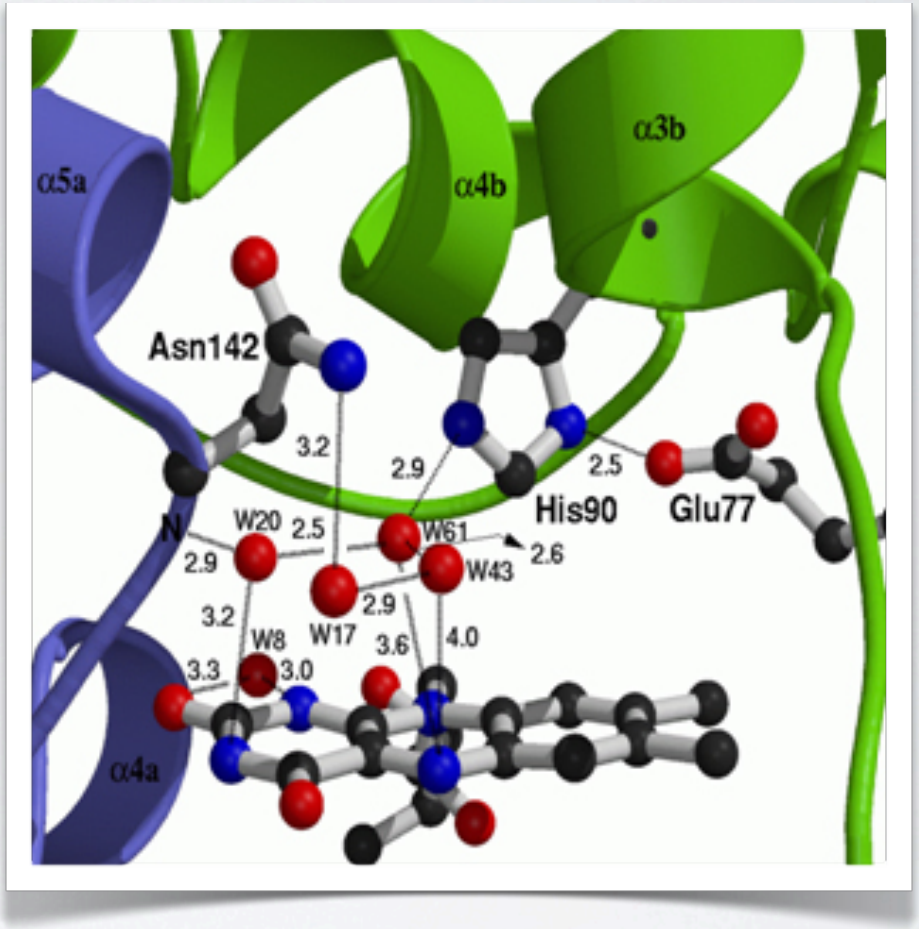
DNA
(NDB)



Small Molecules
(CCDB)

Motivation 1:
Detailed understanding of molecular interactions

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

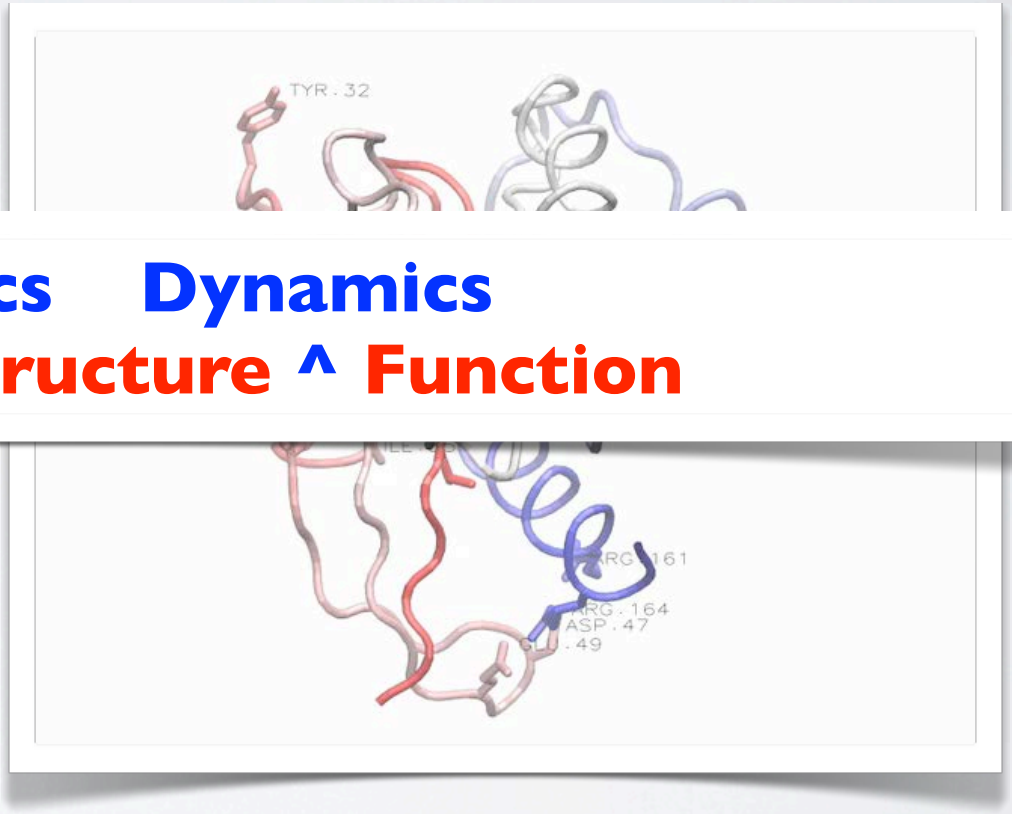


Motivation 1:

Energetics Dynamics

Sequence ^ Structure ^ Function

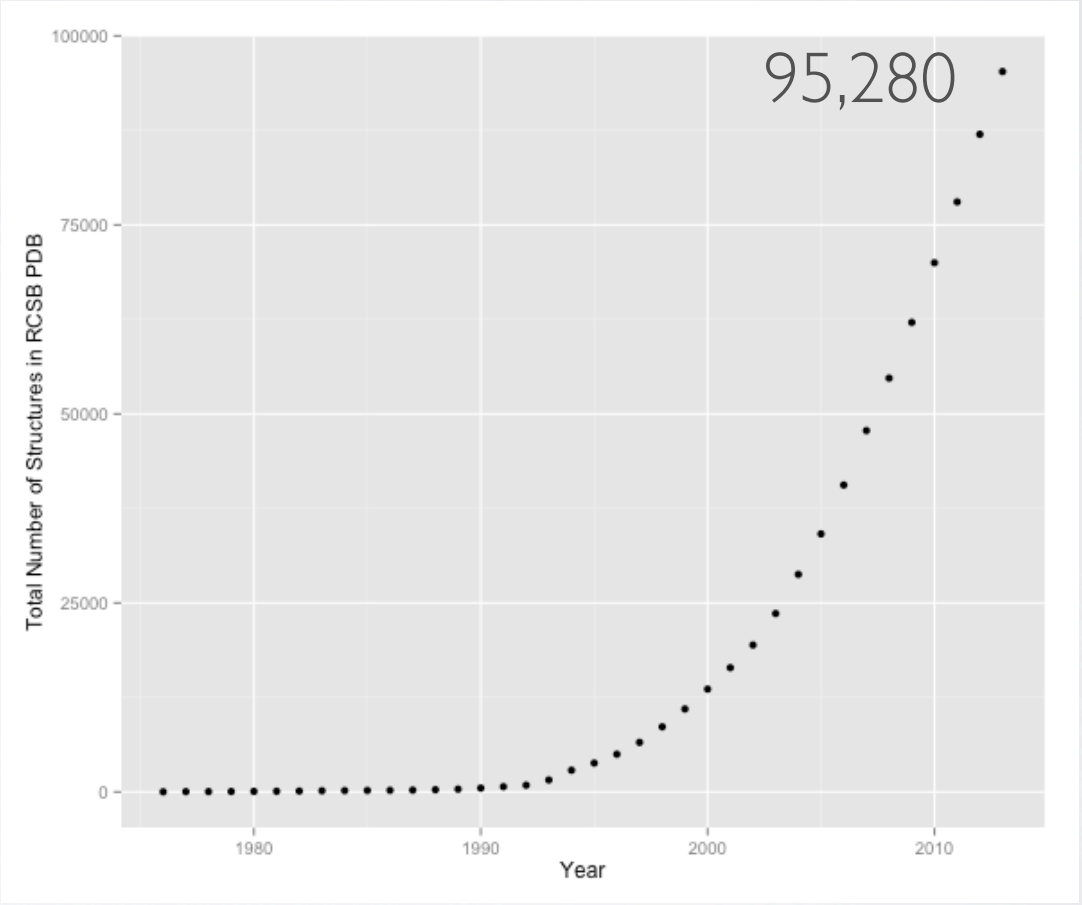
Computational modeling can provide detailed insight into functional interactions, their regulation and potential consequences of perturbation.



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

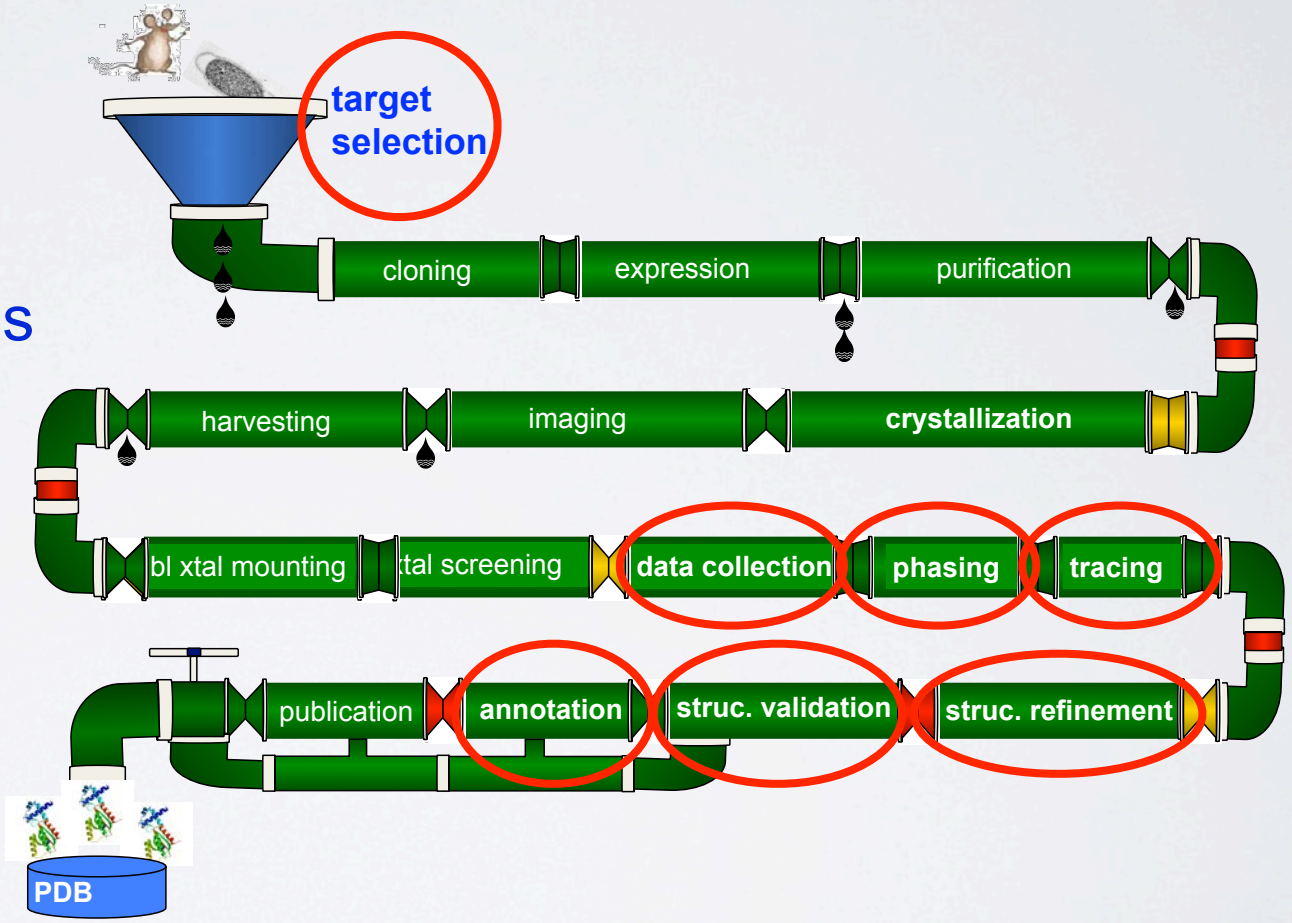
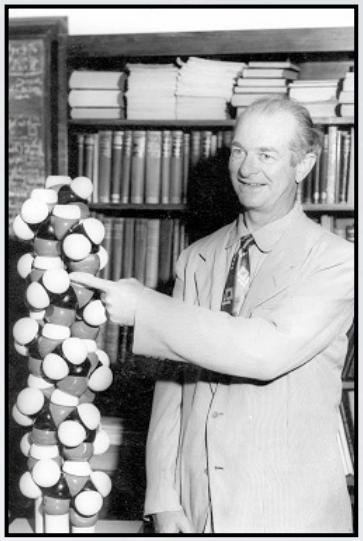


Image Credit: "Structure determination assembly line" Adam Godzik

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

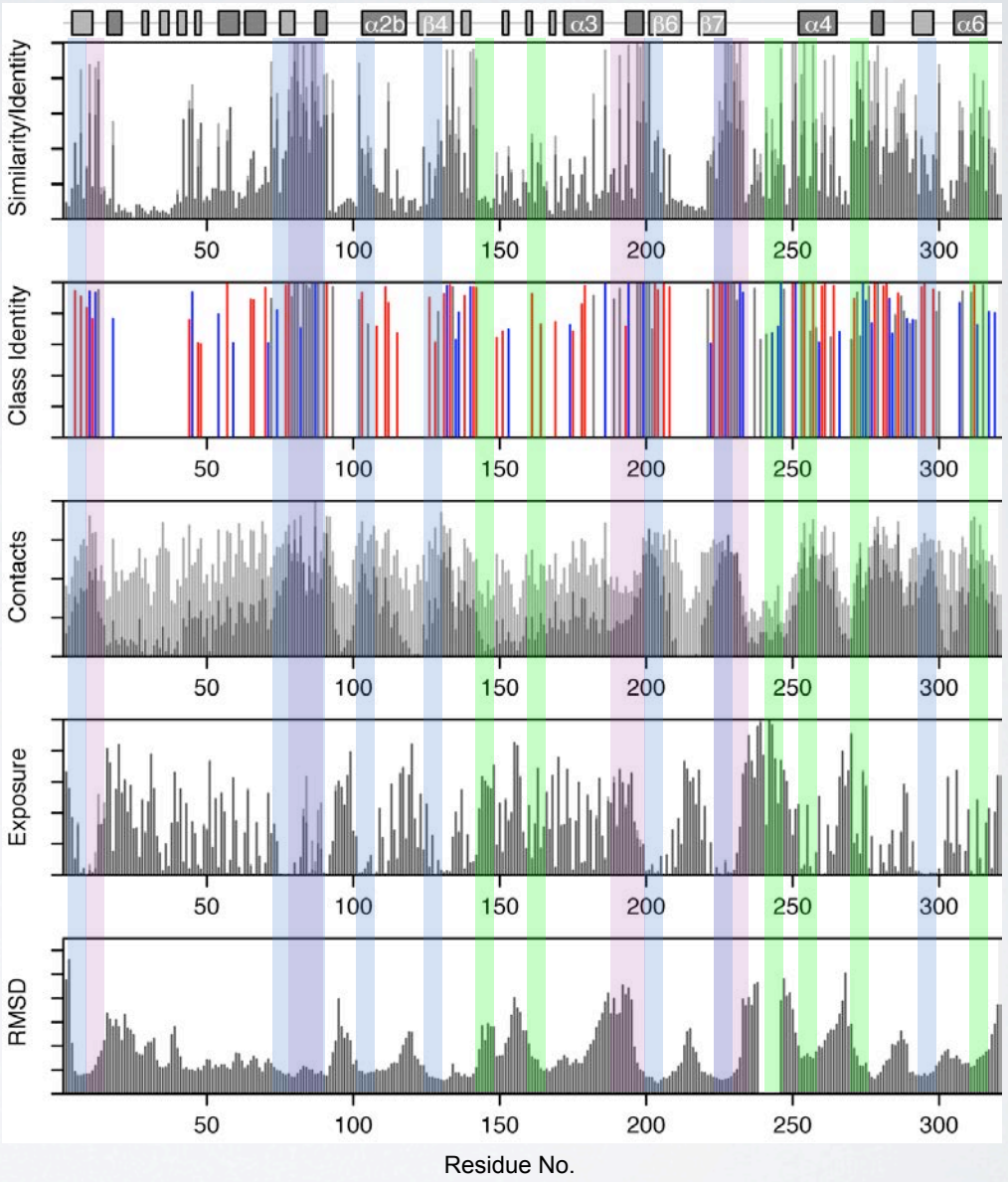
- Structure allows identification of more distant evolutionary relationships

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)

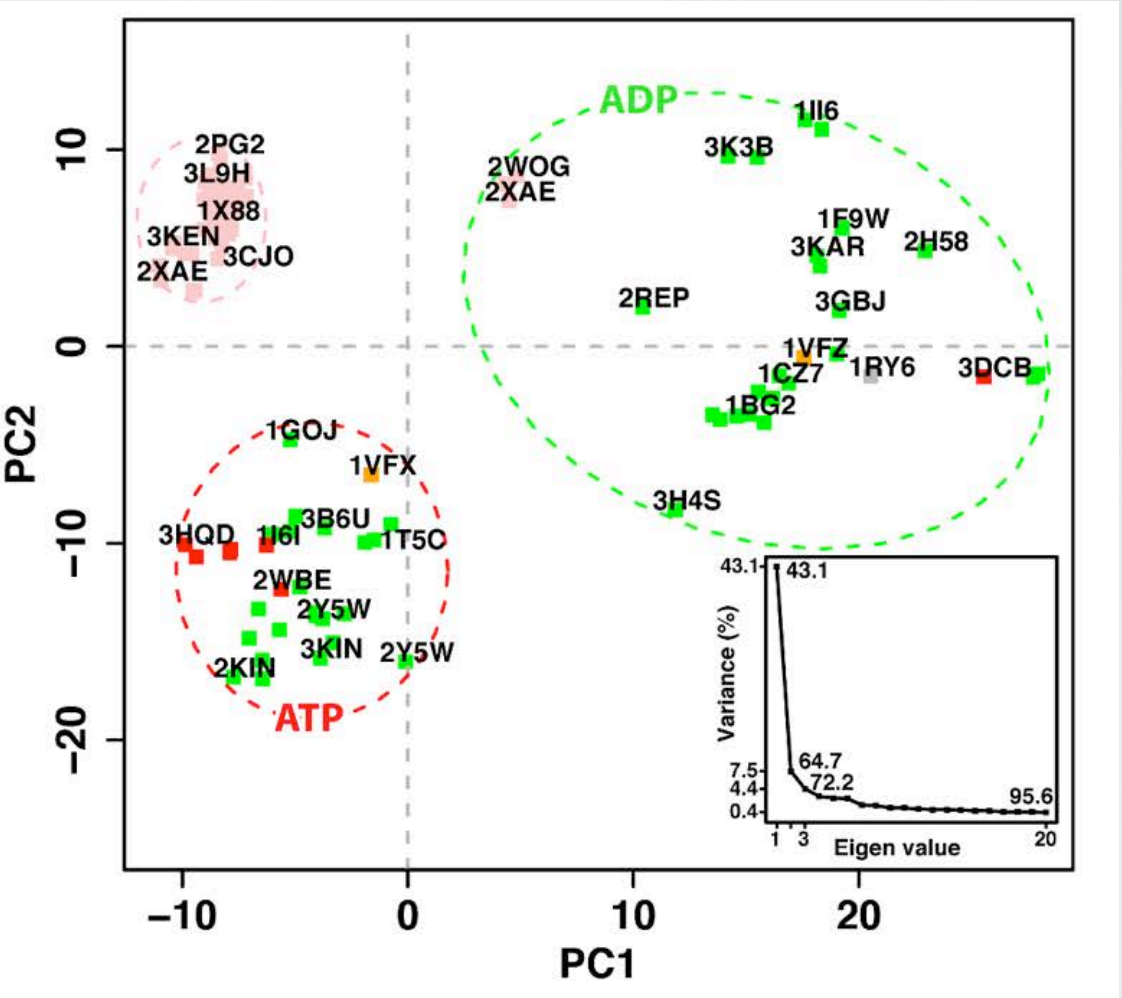
Goals:

- Analysis
- Visualization
- Comparison
- Prediction
- Design

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

Goals:

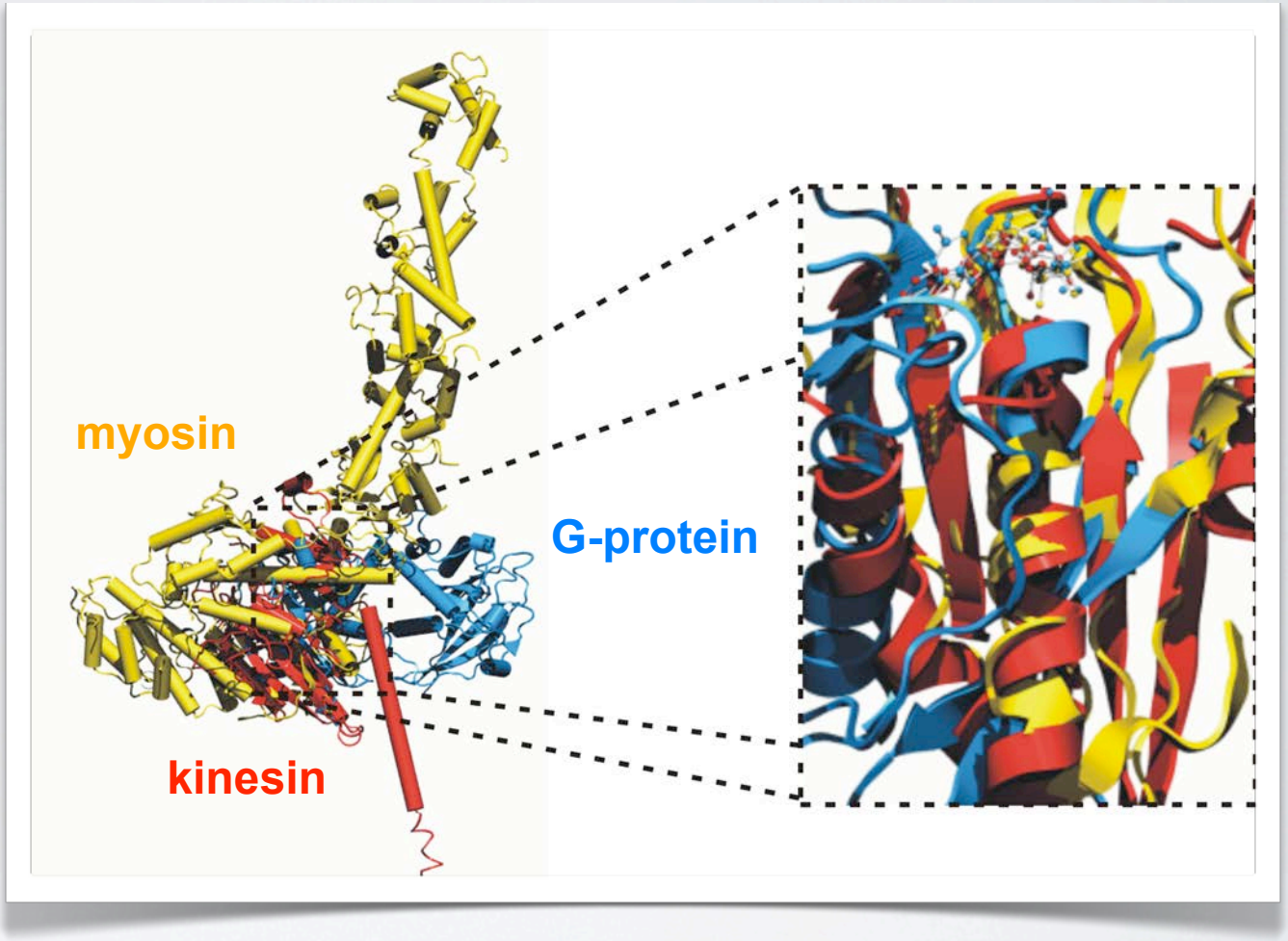
- Analysis
- Visualization
- Comparison
- Prediction
- Design



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

Goals:

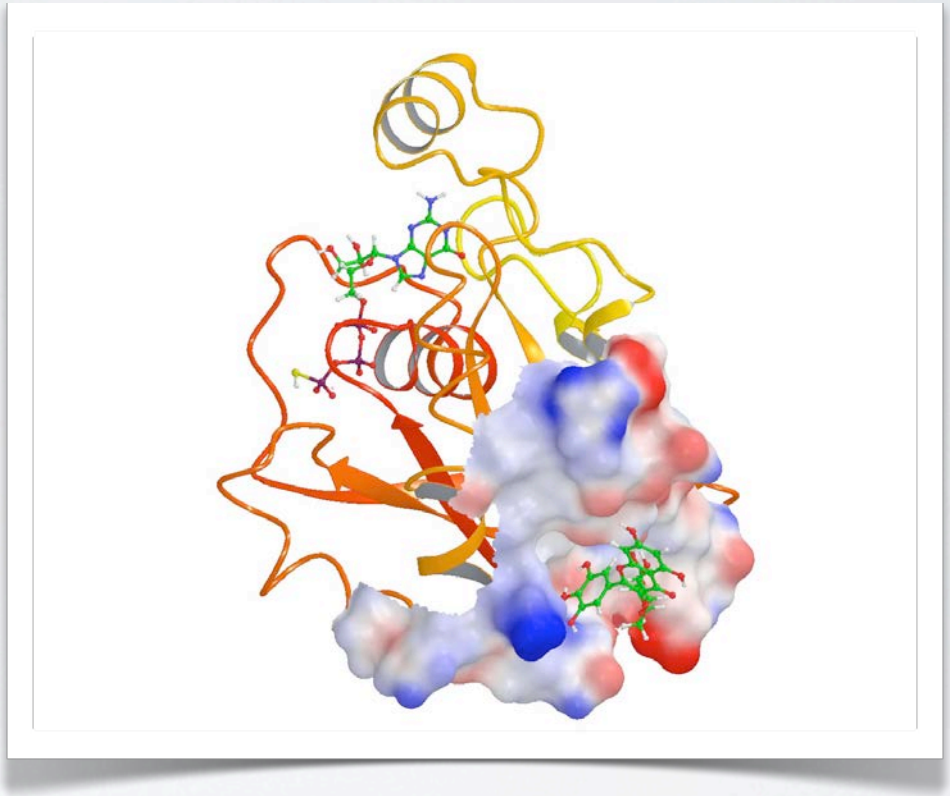
- Analysis
- Visualization
- Comparison
- Prediction
- Design



Grant *et al.* unpublished

Goals:

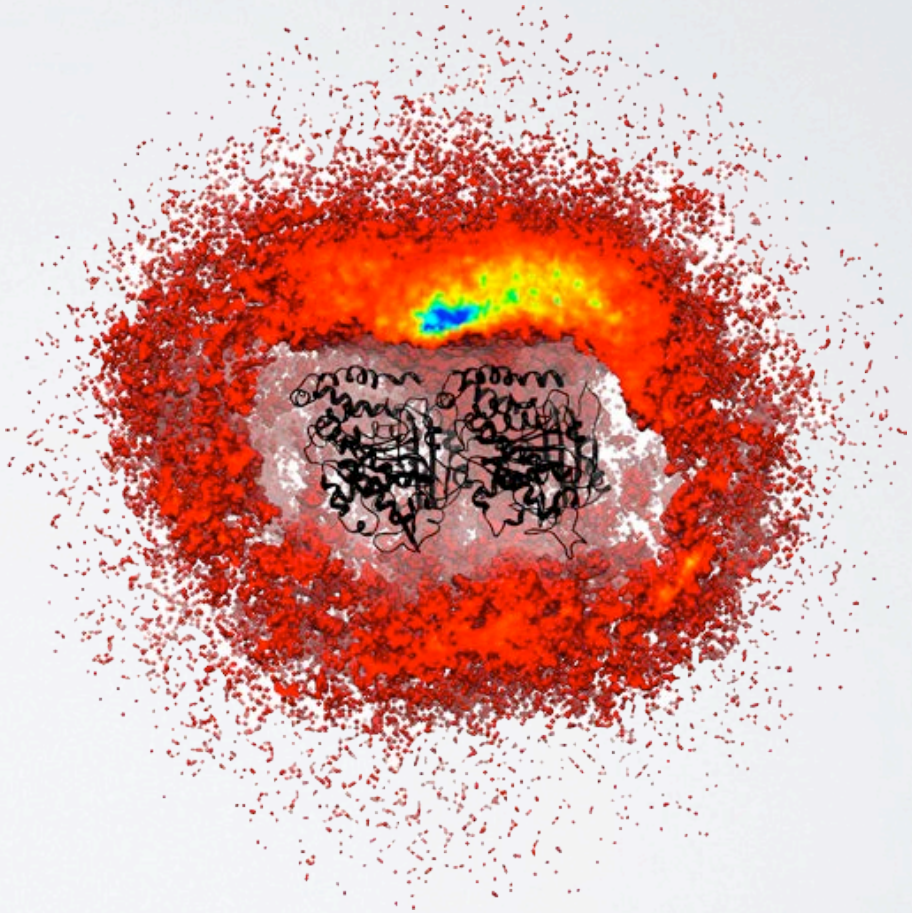
- Analysis
- Visualization
- Comparison
- Prediction
- Design



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

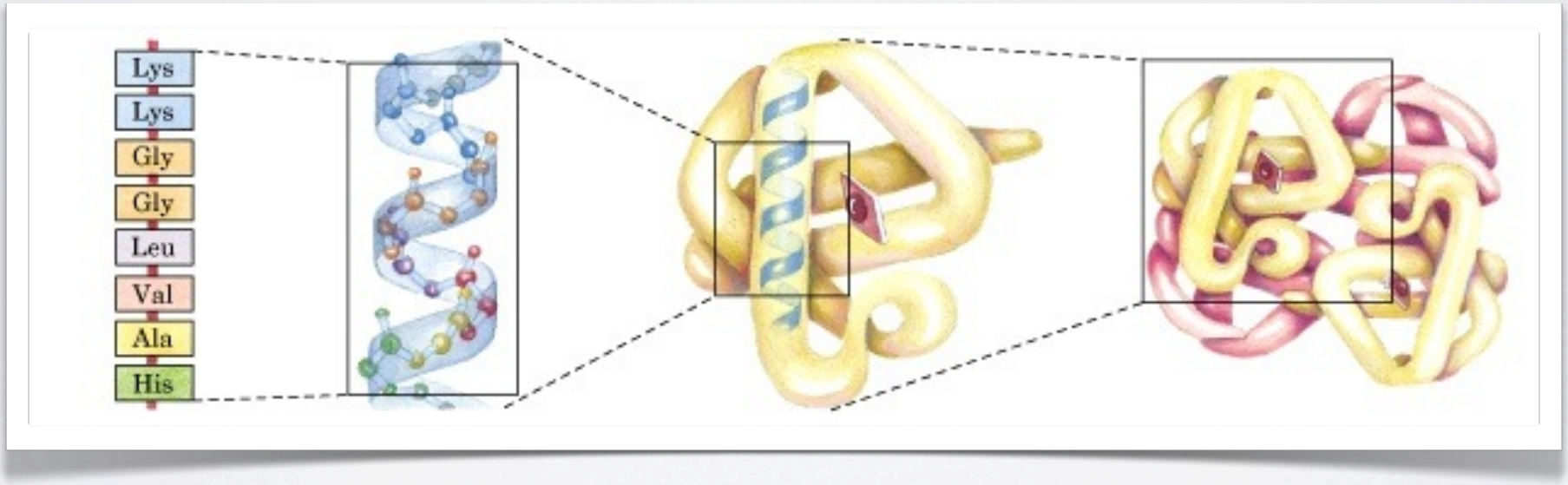
…BREAK…

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

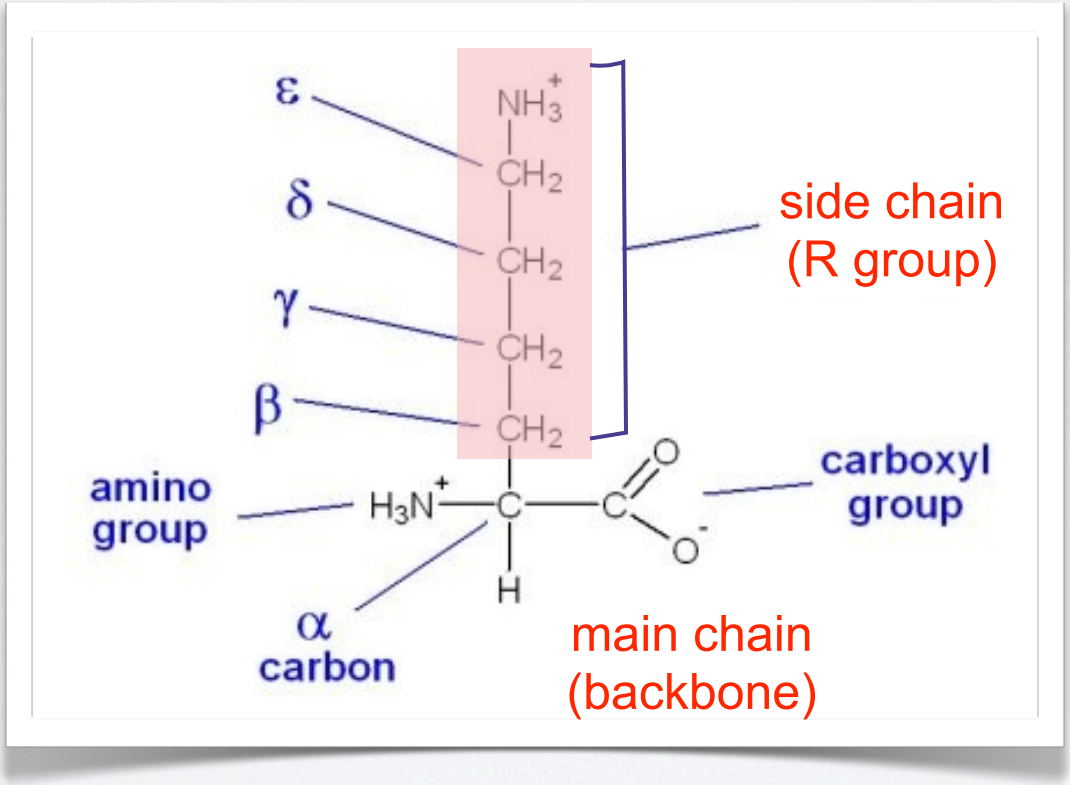
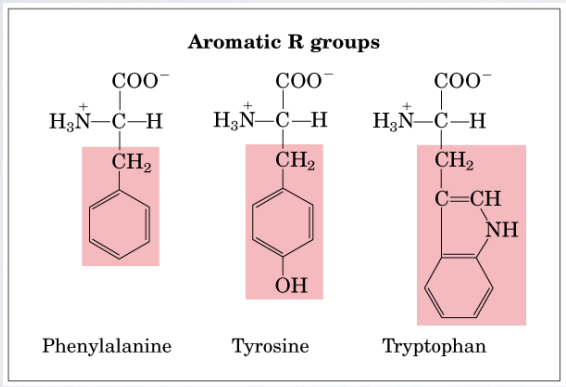
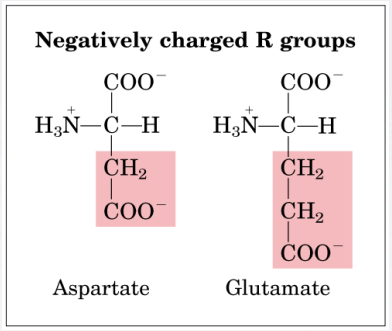
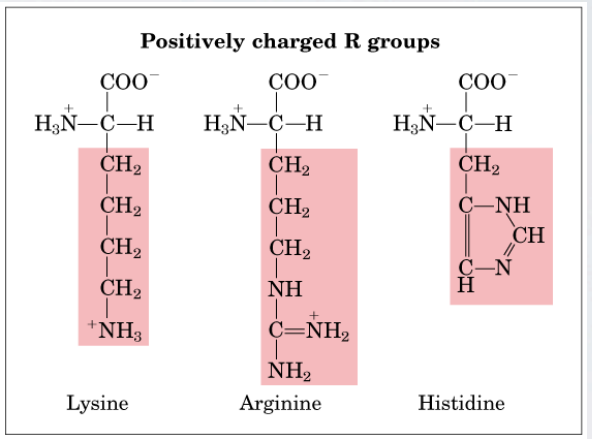


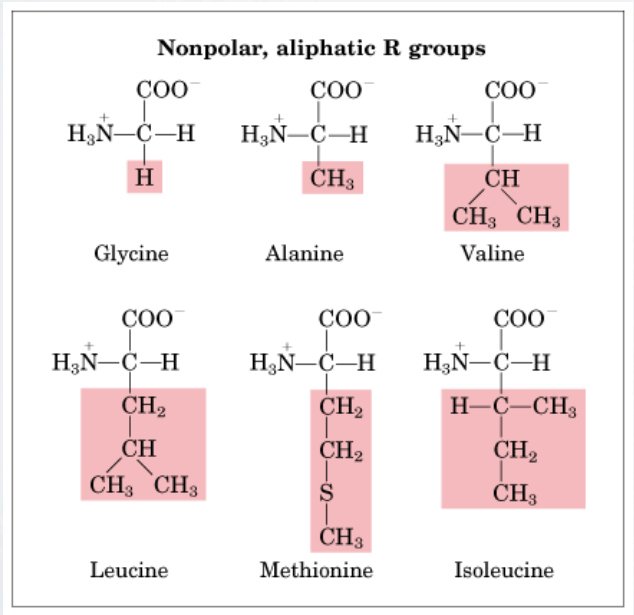
Image from: http://www.ncbi.nlm.nih.gov/books/NBK21581/

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









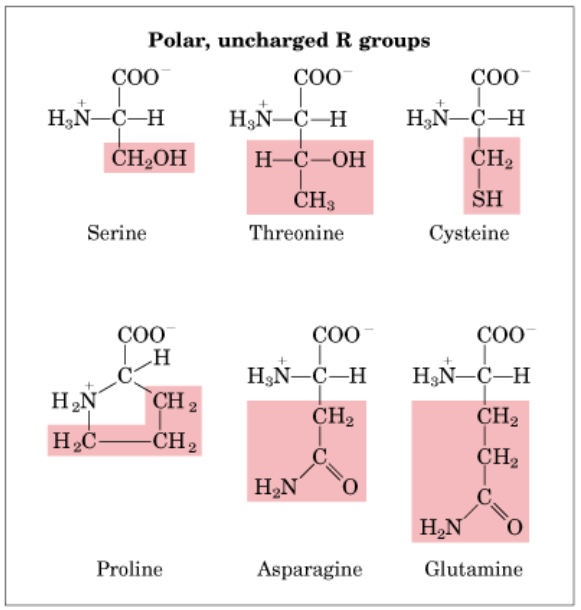
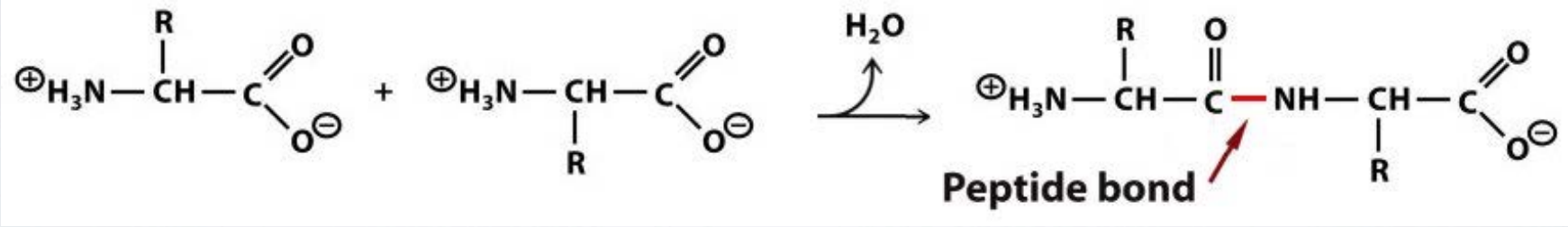


Image from: http://www.ncbi.nlm.nih.gov/books/NBK21581/

AMINO ACIDS POLYMERIZE THROUGH **PEPTIDE BOND** FORMATION



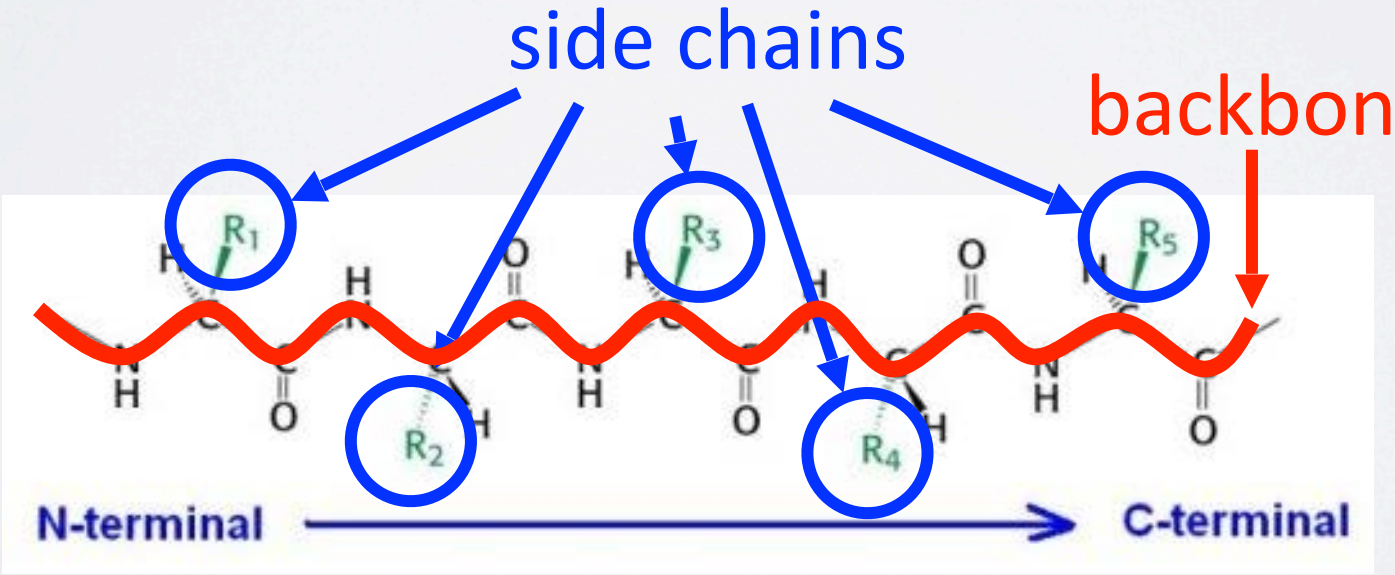
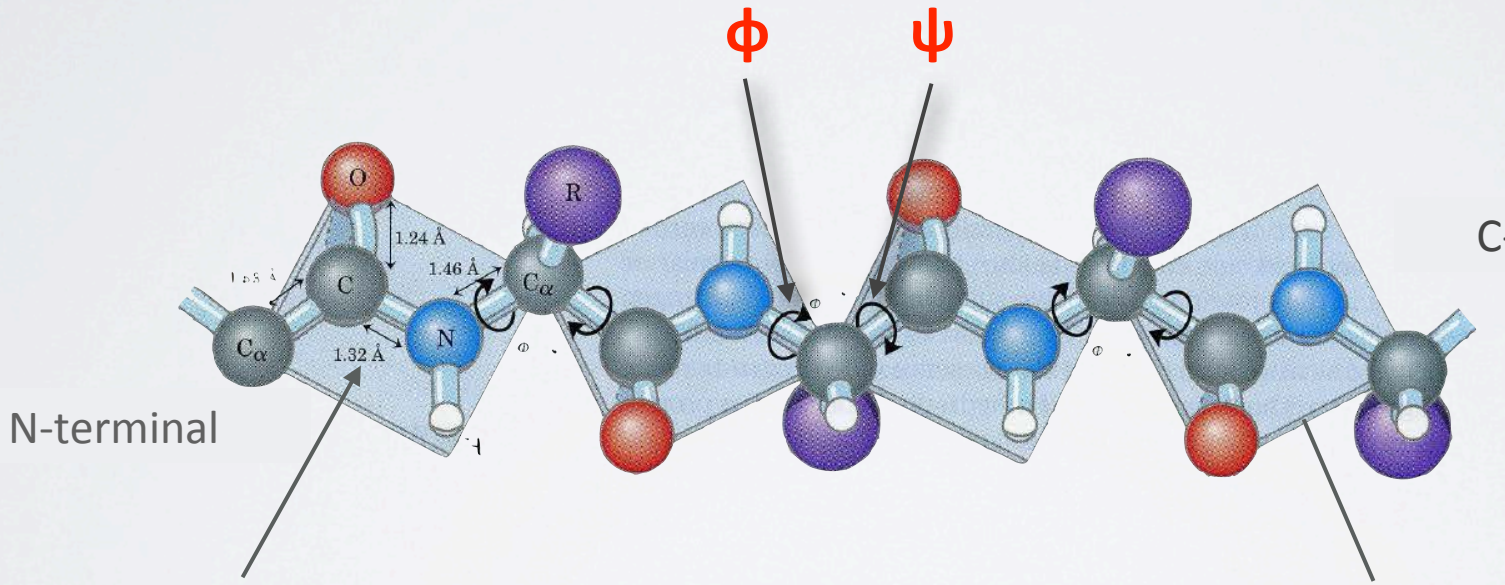


Image from: http://www.ncbi.nlm.nih.gov/books/NBK21581/

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

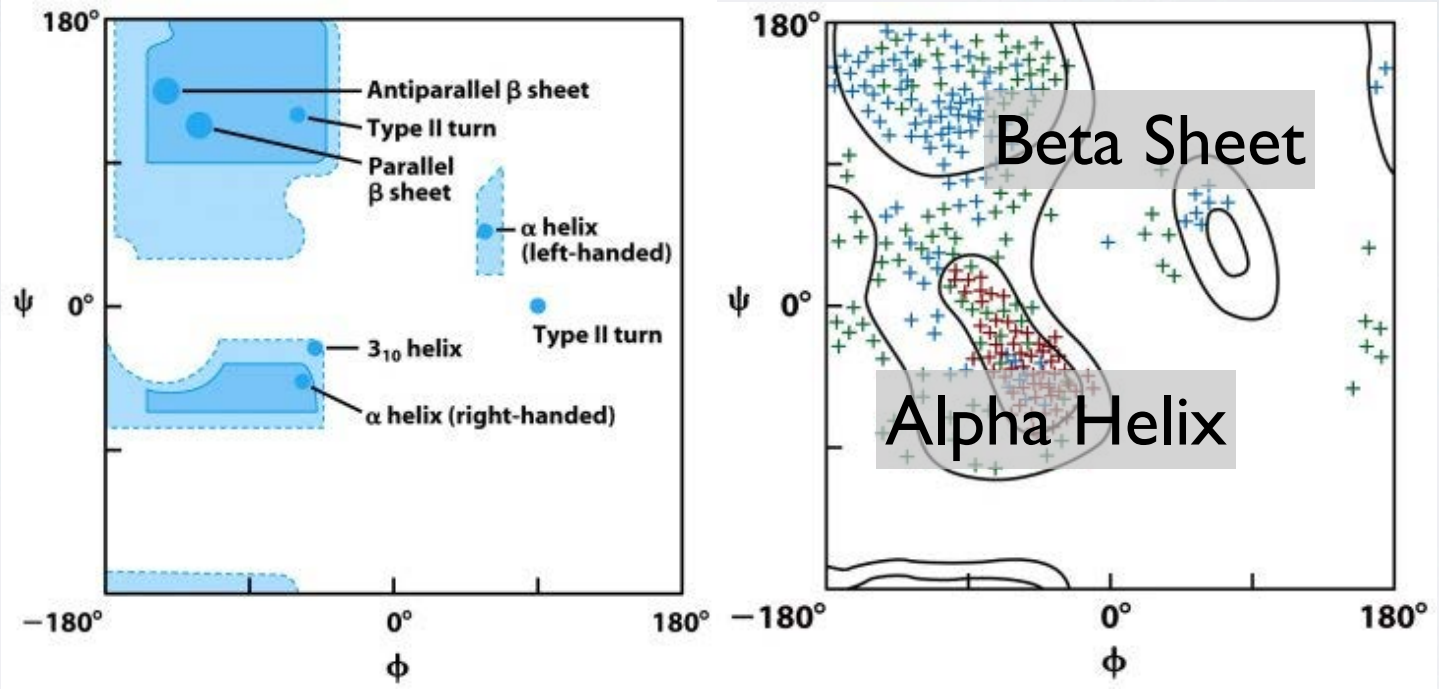


Bond angles and lengths are largely invariant

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/

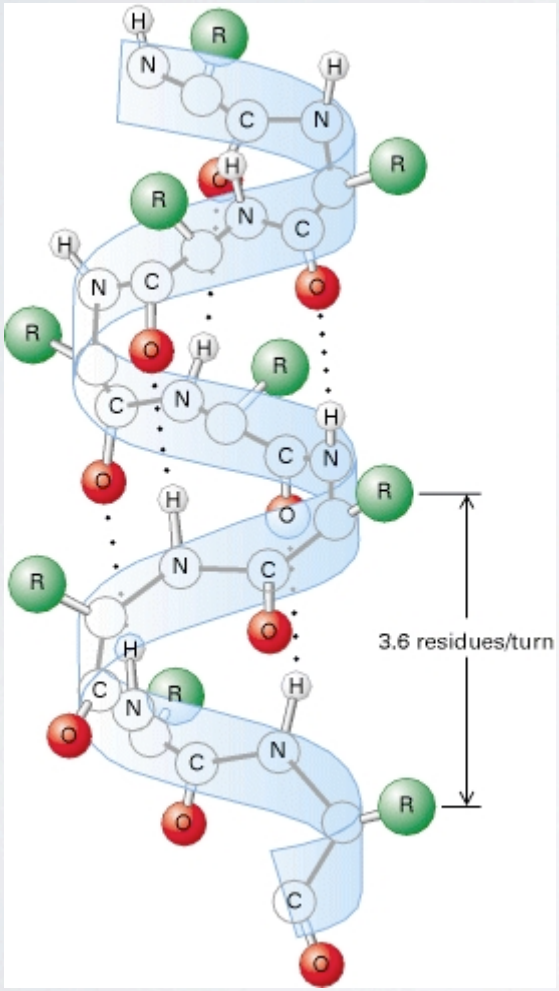
PHI vs PSI PLOTS ARE KNOWN AS **RAMACHANDRAN DIAGRAMS**



- 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



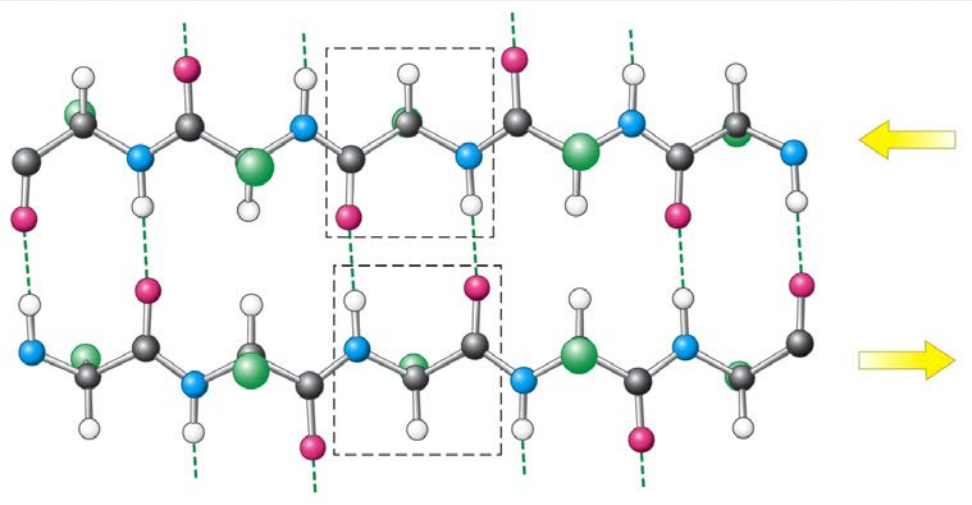
α-helix β-sheets

- Most common from 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 π-helix forms are less common

Hydrogen bond: **i→i+4**

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

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

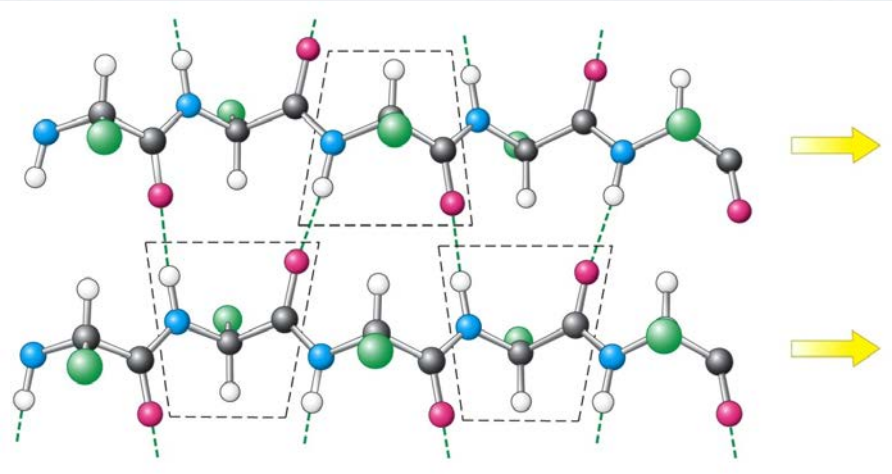


In **antiparallel β-sheets**

- 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**



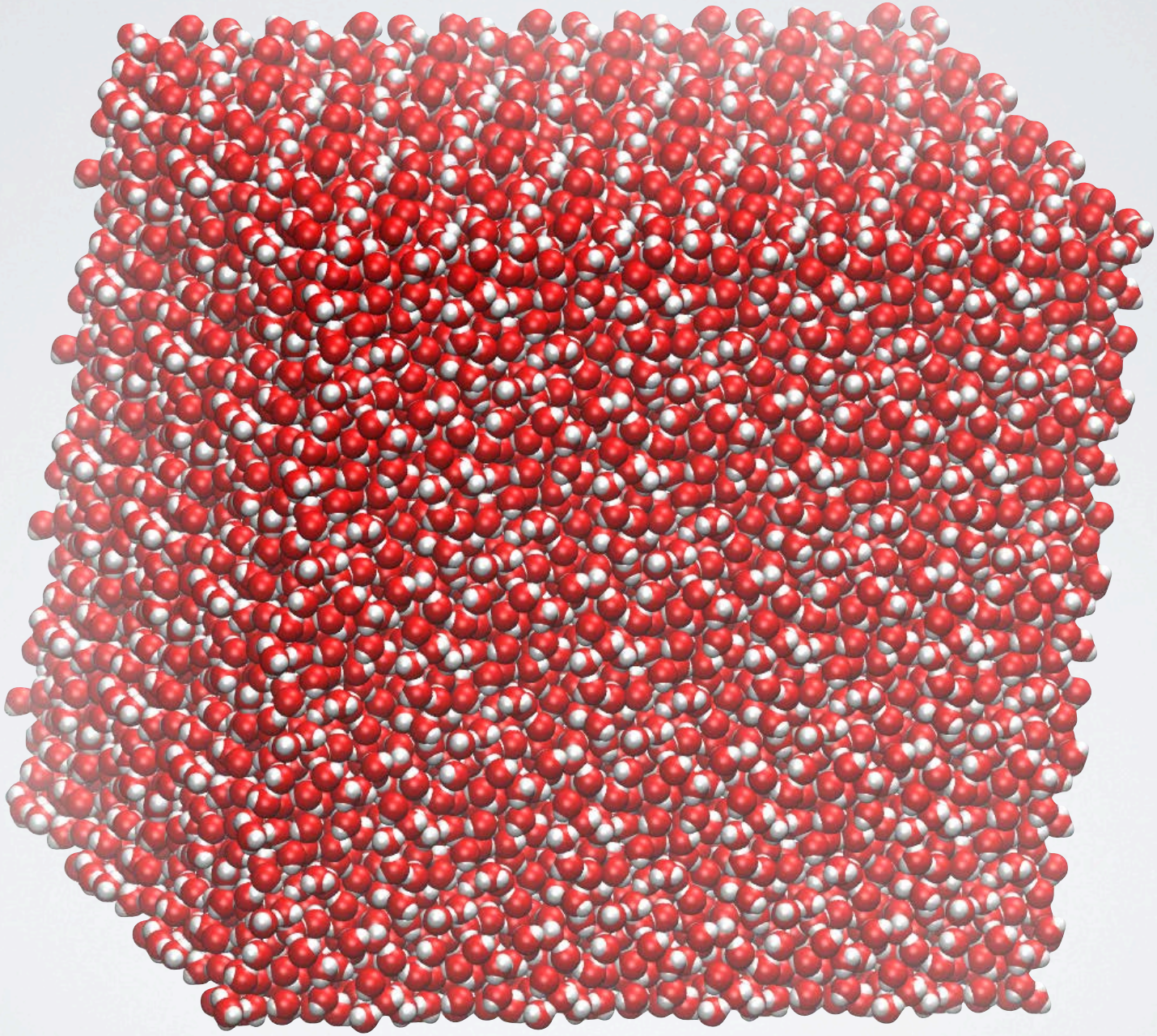
In **parallel β-sheets**

- Adjacent β-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

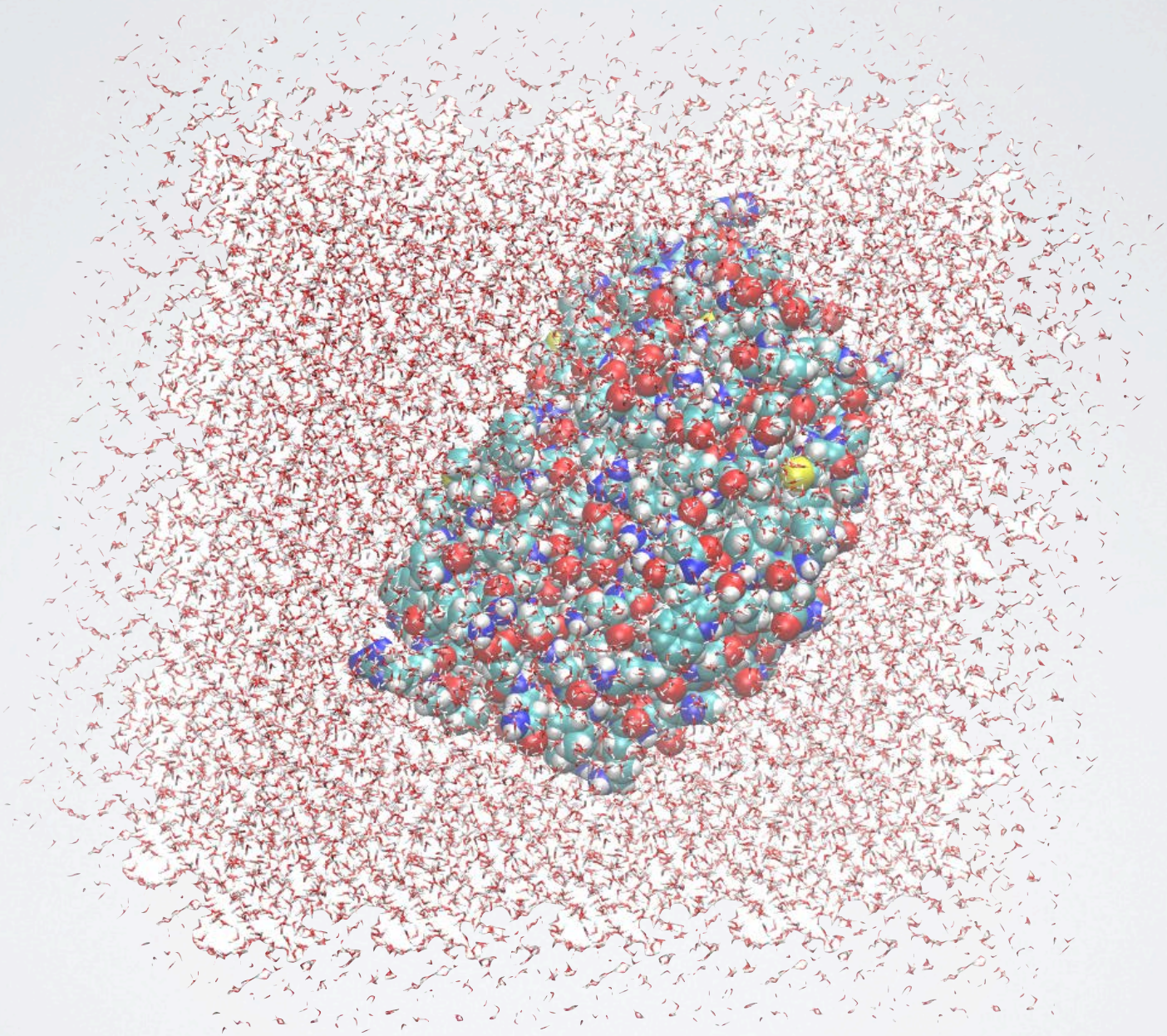
Image from: http://www.ncbi.nlm.nih.gov/books/NBK21581/

WHAT DOES A PROTEIN LOOK LIKE?

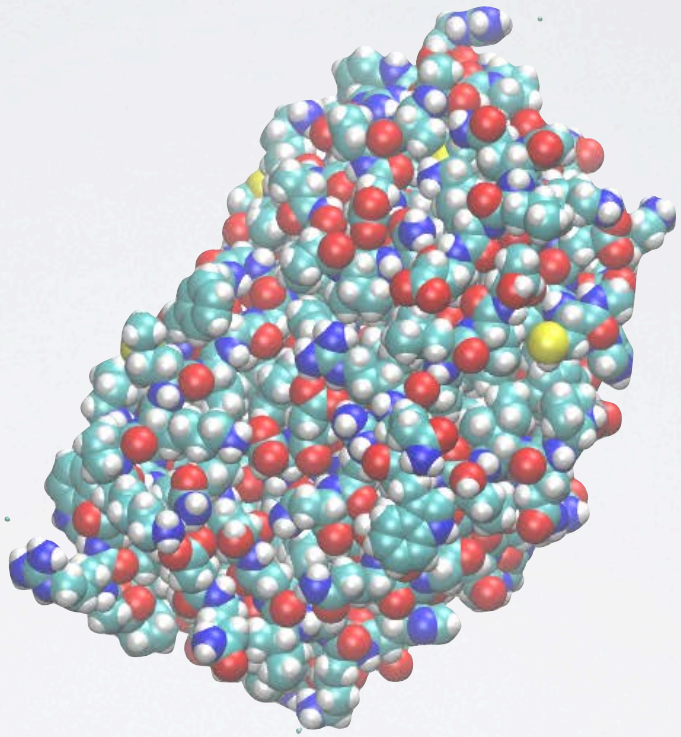
- Hidden in water?
- A close-packed globular object?
- A chain of connected secondary structures?



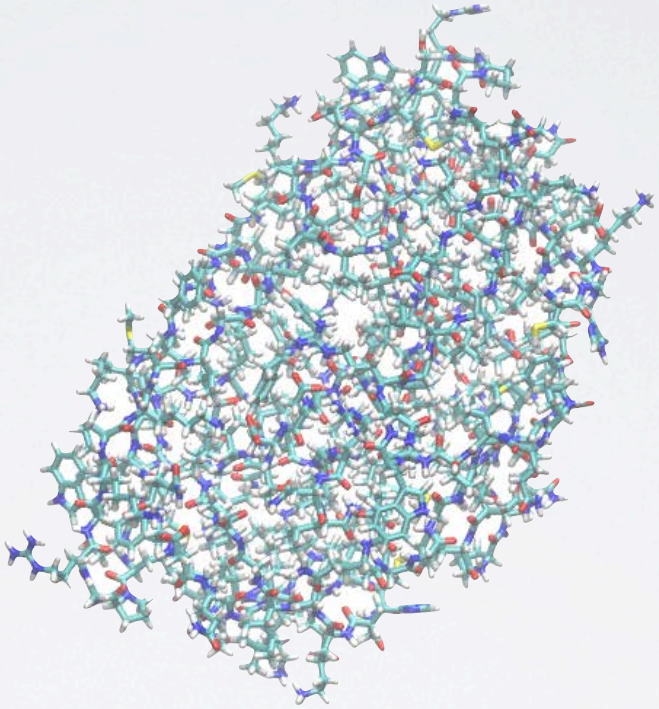
- Proteins are stable in water



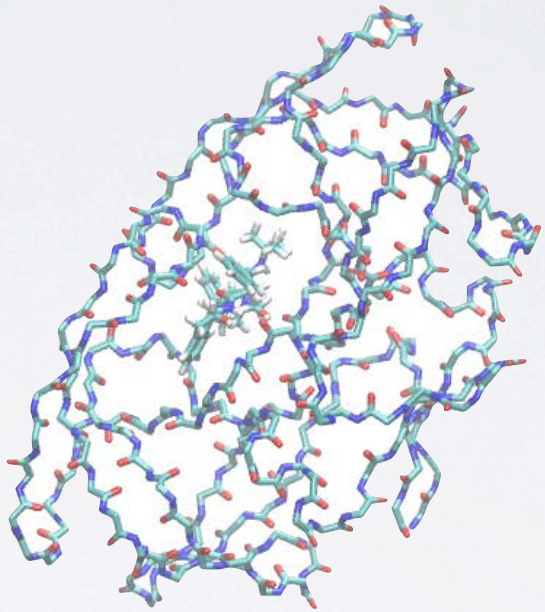
- Proteins closely interact with water



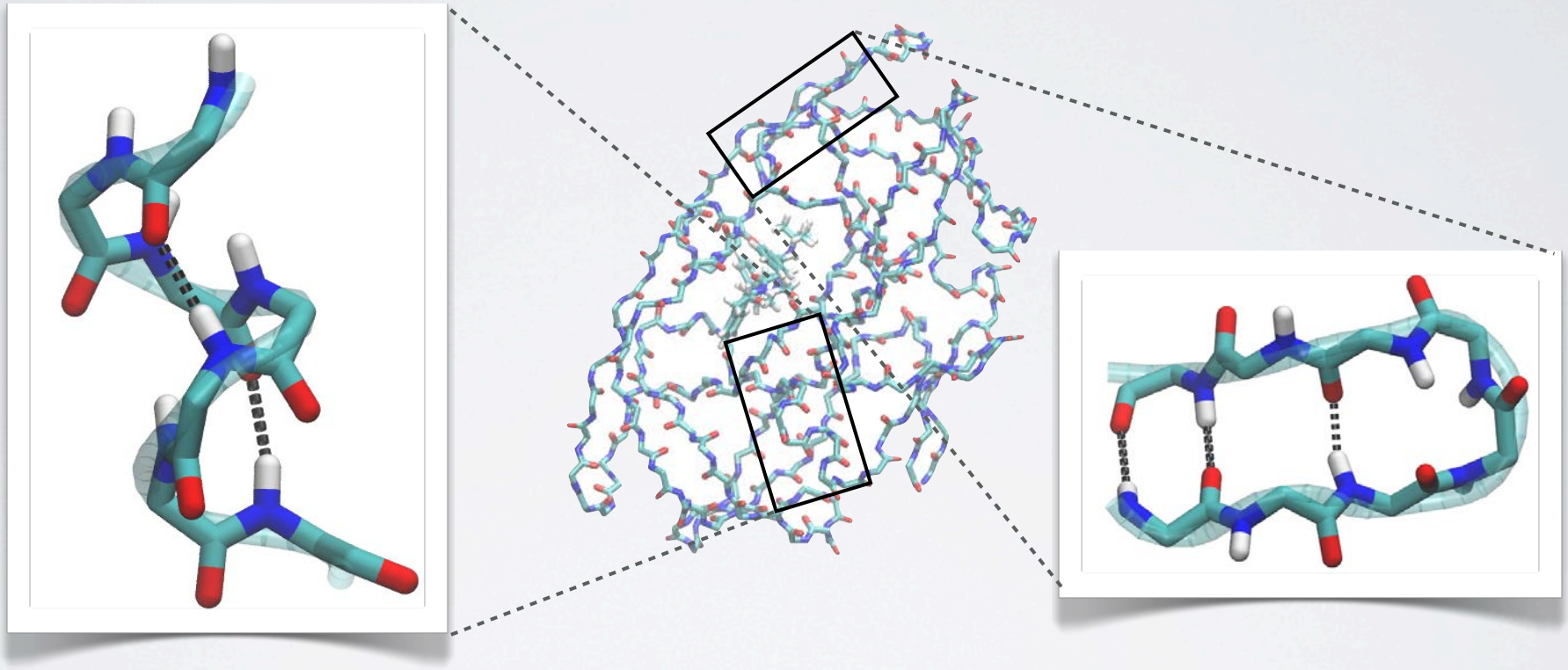
- Proteins are close packed solid but flexible objects



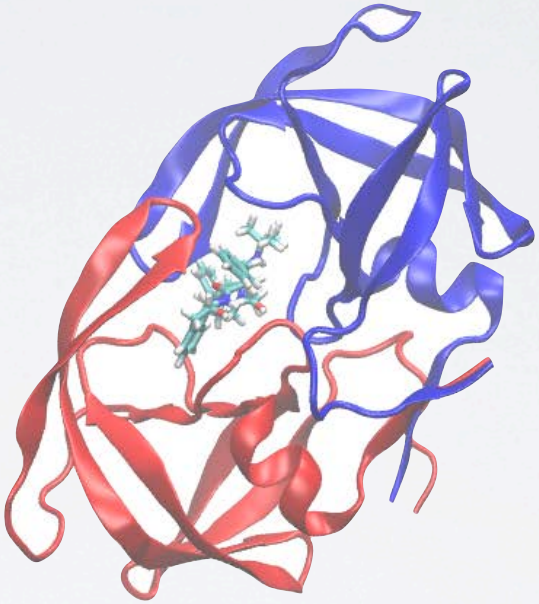
- Due to their large size and complexity it is often hard to see whats important in the structure



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



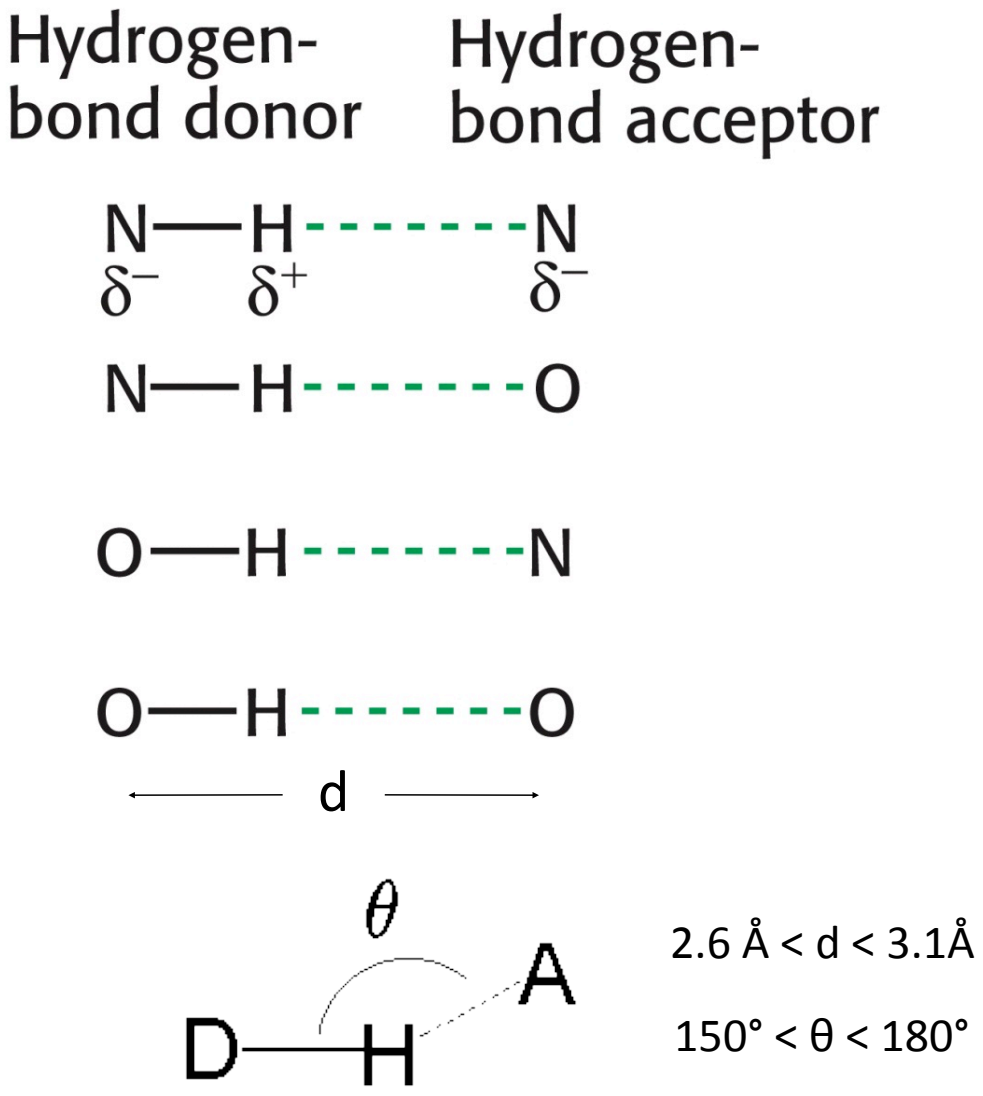
- Backbone or main-chain representation can help trace chain topology & reveal secondary structure



- Simplified secondary structure representations are commonly used
- Now we can clearly see 2º, 3º and 4º structure

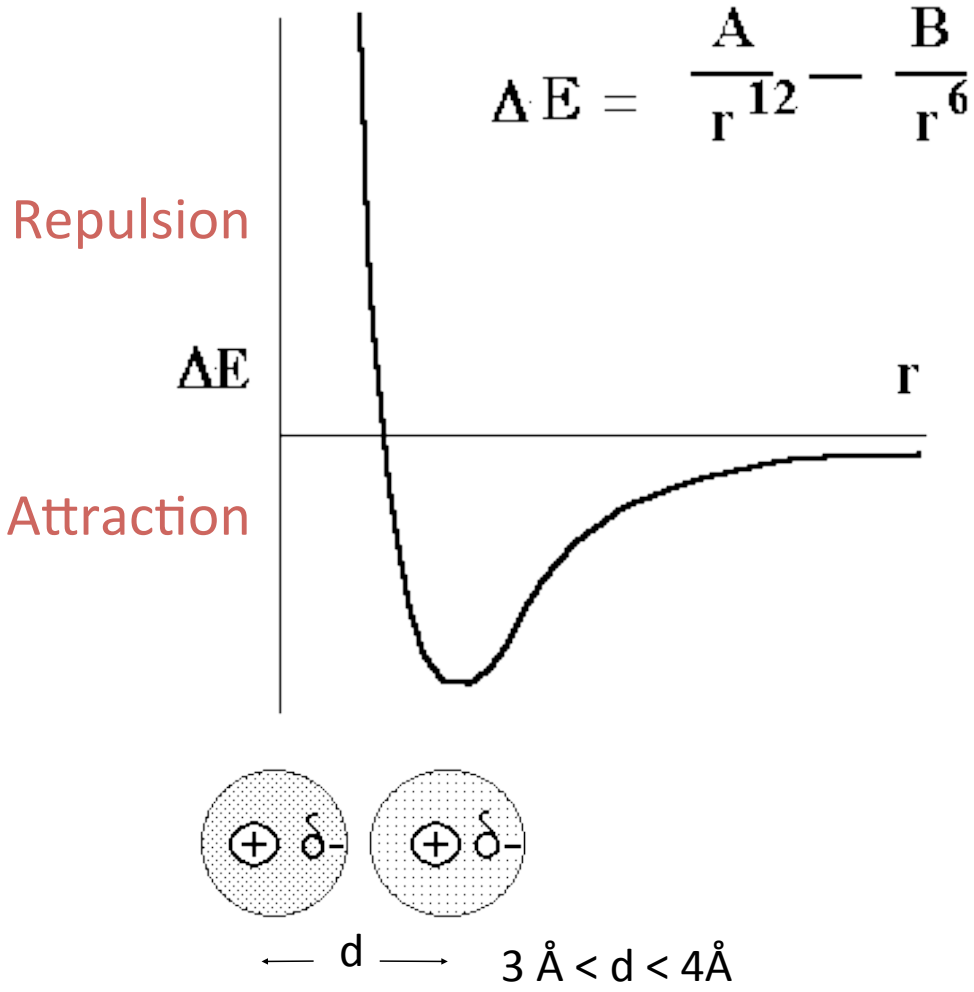
Key forces affecting structure:

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



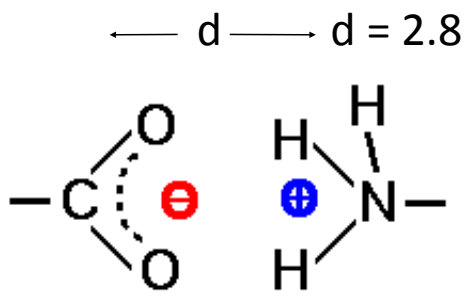
Key forces affecting structure:

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



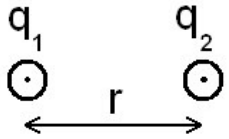
Key forces affecting structure:

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



carboxyl group and amino group

(some time called IONIC BONDs or SALT BRIDGEs)



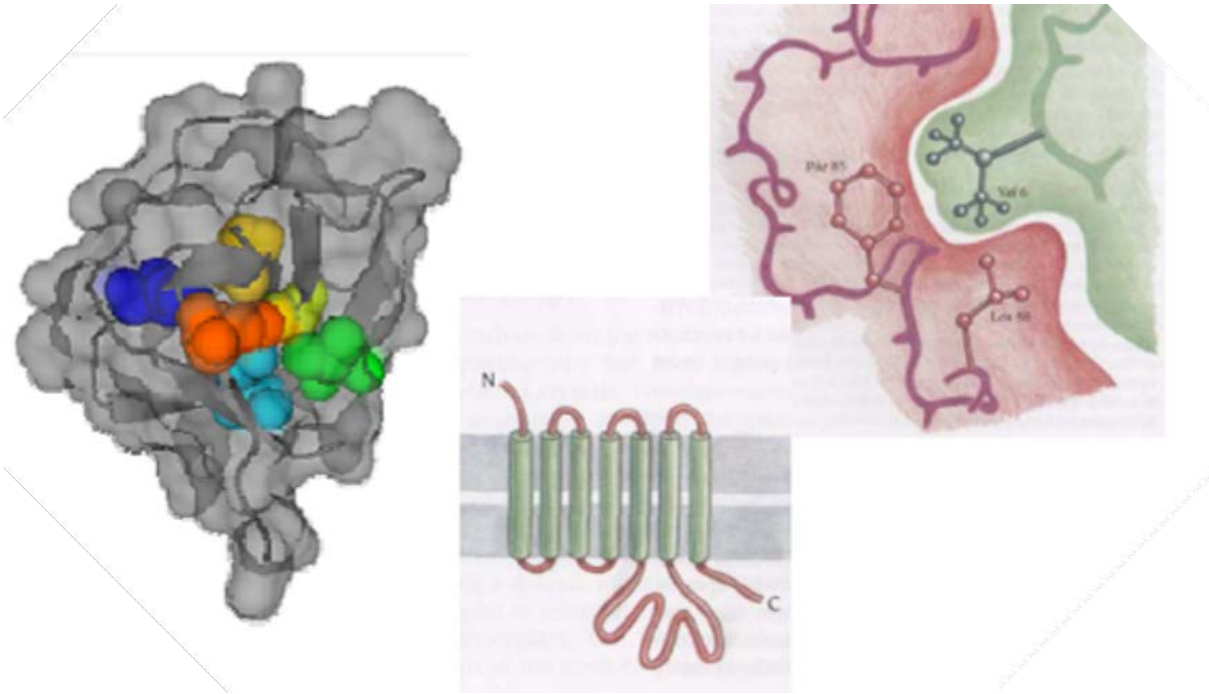
Coulomb's law

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

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

Key forces affecting structure:

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

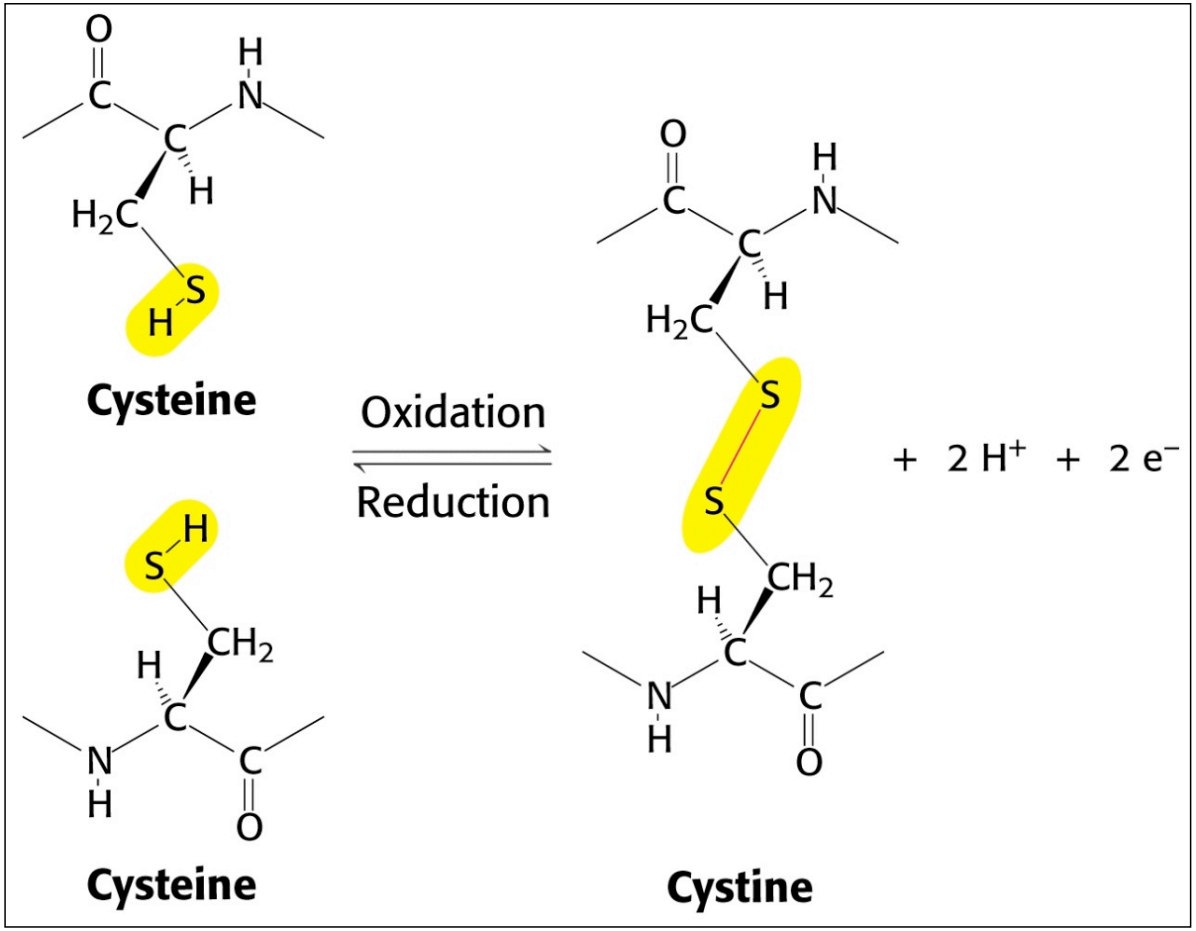


The force that causes hydrophobic molecules or nonpolar portions of molecules to aggregate together rather than to dissolve in water is called Hydrophobicity (*Greek, "water fearing"*). This is not a separate bonding force; rather, it is the result of the energy required to insert a nonpolar molecule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

RCSB 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

Customize This Page

Available on the App Store

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

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

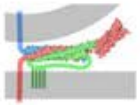
Biological Macromolecular Resource

Full Description

Learn: Featured Molecules Hide

Structural View of Biology — List View of Archive By: Title | Date | Category

PDB-101 — Protein Synthesis

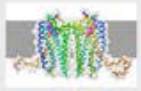


Molecule of the Month

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.

Full Article



Protein Structure Initiative Featured System

G Proteins and Cancer

Miscommunication is the hallmark of cancer. Normally, our cells are in constant communication, deciding how to share resources, determining the best time to grow, and if necessary, the best time to die. Cancer cells, on the other hand, typically have corrupted these lines of communication, allowing them to grow without limits and selfishly steal resources for themselves. GPCRs (G-protein-coupled receptors) are among the many different molecules of communication that are changed when a normal cell is transformed into a cancer cell.

Full Article | Archive | PSI Structural Biology Knowledgebase

Explore Archive Hide

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

Show all

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)

Latest Structures Hide

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

New Features Hide

Latest release: September 2013

Visualize, browse & search symmetry / stoichiometry

Website Release Archive:

RCSB PDB News Hide

Weekly | Quarterly | Yearly

2013-11-12

Upcoming Meeting: ABRCMS

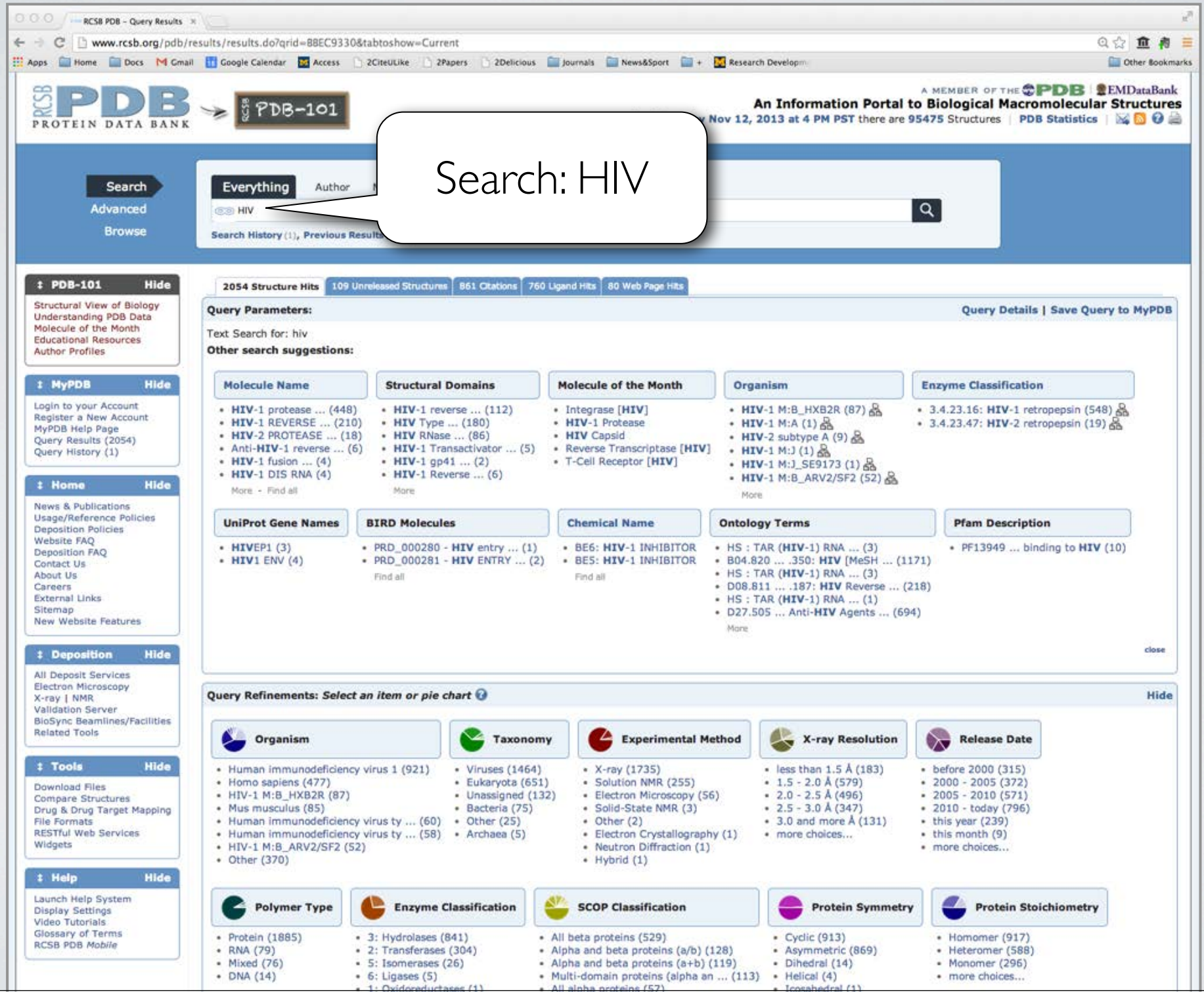
Meet the RCSB PDB at booth 715 at the Annual Biomedical Research Conference for Minority Students (Nov 13-16; Nashville, TN). more

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

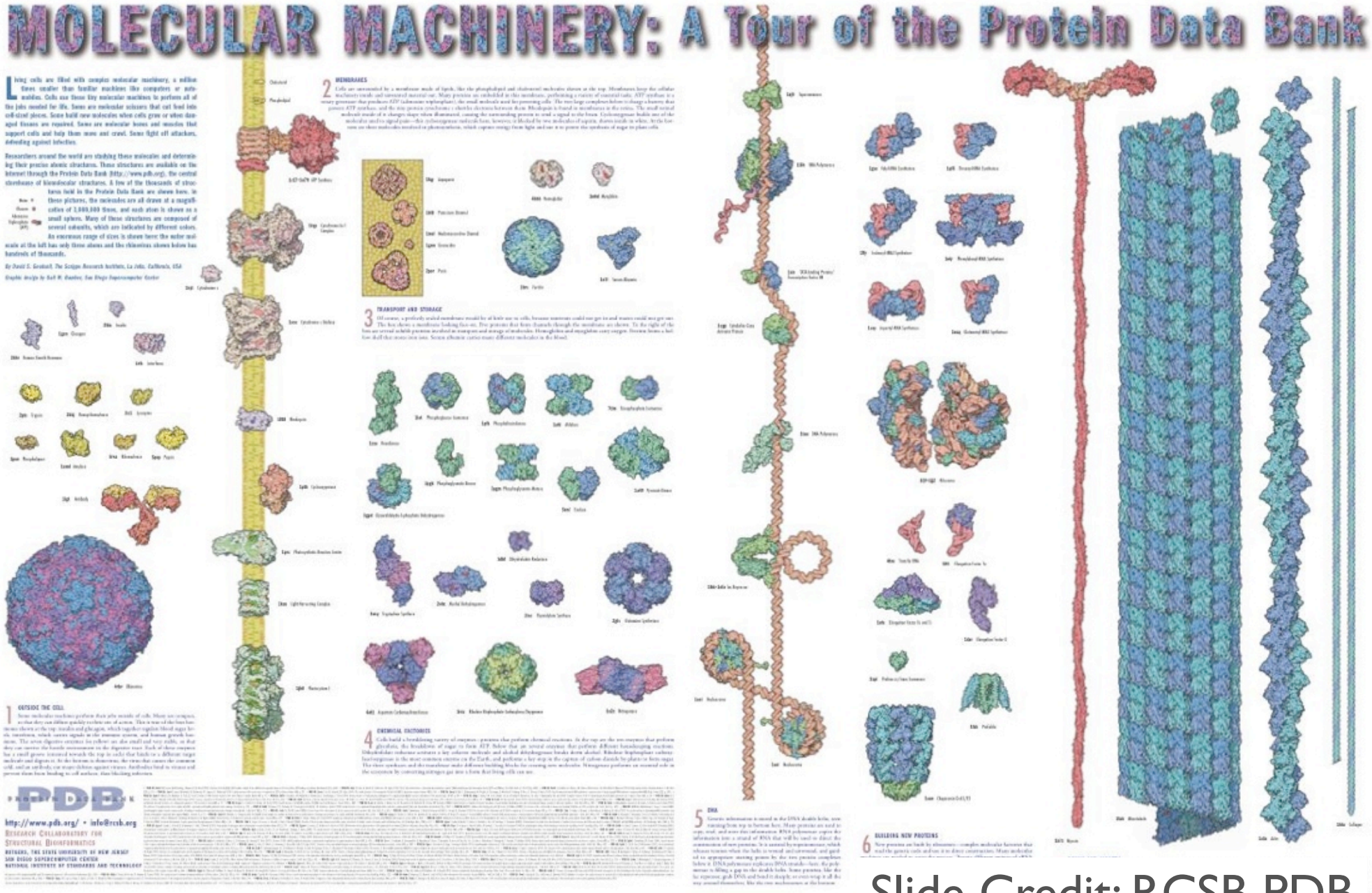
New Structures Hide

Latest Release

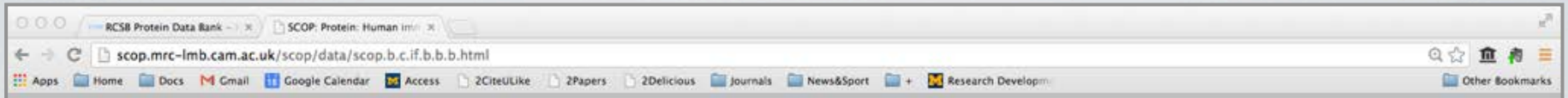
New Structure Papers


RCSB PDB - Query Results
www.rcsb.org/pdb/results/results.do?qrid=88EC9330&tabtoshow=Current
PROTEIN DATA BANK
PDB-101
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
Search: HIV
Search
Advanced
Browse
Everything
Author
HIV
Search History (1), Previous Results
2054 Structure Hits
109 Unreleased Structures
861 Citations
760 Ligand Hits
80 Web Page Hits
Query Parameters:
Query Details | Save Query to MyPDB
Text Search for: hiv
Other search suggestions:
Molecule Name
Structural Domains
Molecule of the Month
Organism
Enzyme Classification
UniProt Gene Names
BIRD Molecules
Chemical Name
Ontology Terms
Pfam Description
Query Refinements: Select an item or pie chart
Organism
Taxonomy
Experimental Method
X-ray Resolution
Release Date
Polymer Type
Enzyme Classification
SCOP Classification
Protein Symmetry
Protein Stoichiometry


Search: 1HSG
(PDB ID)
RCSB Protein Data Bank
www.rcsb.org/pdb/explore/explore.do?structureId=1HSG
PROTEIN DATA BANK
PDB-101
An Information Portal to Biological Macromolecular Structures
Search
Advanced
Browse
Everything
Author
1HSG
Search History (1), Previous Results (2054)
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 hydroxyaminopentane 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
To save personal annotations, please login to your MyPDB account.
Deposition Summary
Authors: Chen, Z.
Deposition: 1995-03-31
Release: 1996-04-03
PDB-101
Structural View of Biology
Understanding PDB Data
Molecule of the Month
Educational Resources
Author Profiles
MyPDB
Login to your Account
Register a New Account
MyPDB Help Page
Query Results (2054)
Query History (1)
Home
News & Publications
Usage/Reference Policies
Deposition Policies
Website FAQ
Deposition FAQ
Contact Us
About Us
Careers
External Links
Sitemap
New Website Features
Deposition
All Deposit Services
Electron Microscopy
X-ray | NMR
Validation Server
BioSync Beamlines/Facilities
Related Tools
Tools
Download Files
Compare Structures
Drug & Drug Target Mapping
File Formats
RESTful Web Services
Widgets
Help
Launch Help System
Display Settings
Video Tutorials
Glossary of Terms
RCSB PDB Mobile



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

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.

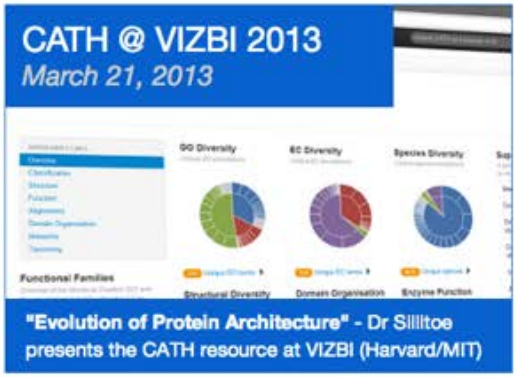
Searching CATH

- 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

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

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, Dessailly BH, Dawson NL, Furnham N, Lee D, Lees JG, Lewis TE, Studer RA, Rentzsch R, Yeats C, Thornton JM, Orengo CA

Nucleic Acids Res. 2013 Jan Pubmed: 23203873

RCSB Protein Data Bank | SCOP: Protein: Human | CATH Superfamily 2.40.70

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

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

CATH | Home | Search | Browse | Download | About | Support | 1hsg

CATH Superfamily 2.40.70.10

Acid Proteases

Home / Superfamily 2.40.70.10

SUPERFAMILY LINKS

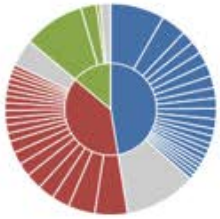
- **Summary**
- Superfamily Superposition
- Classification / Domains
- Alignments
- Structural Neighbourhood
- Functional Annotations
- Taxonomy Browser
- Multi-Domain Organisation

Functional Families

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

GO Diversity

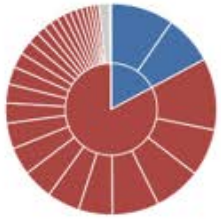
Unique GO annotations



111 Unique GO terms >

EC Diversity

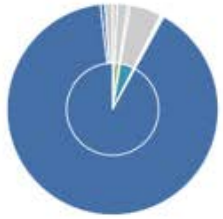
Unique EC annotations



36 Unique EC terms >

Species Diversity

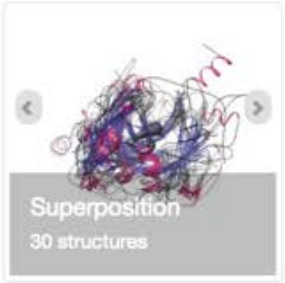
Unique species annotations



1468 Unique species >

Structural Diversity

Structural domains within this superfamily



Domain Organisation

View multi-domain architectures via ArchSchema (Laskowski/EBI)

ArchSchema (requires Java)

Enzyme Function

Evolution of Enzyme Function via FunTree (Furnham/EBI)

FunTree (opens new window)

Superfamily Summary

A general summary of information for this superfamily.

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**

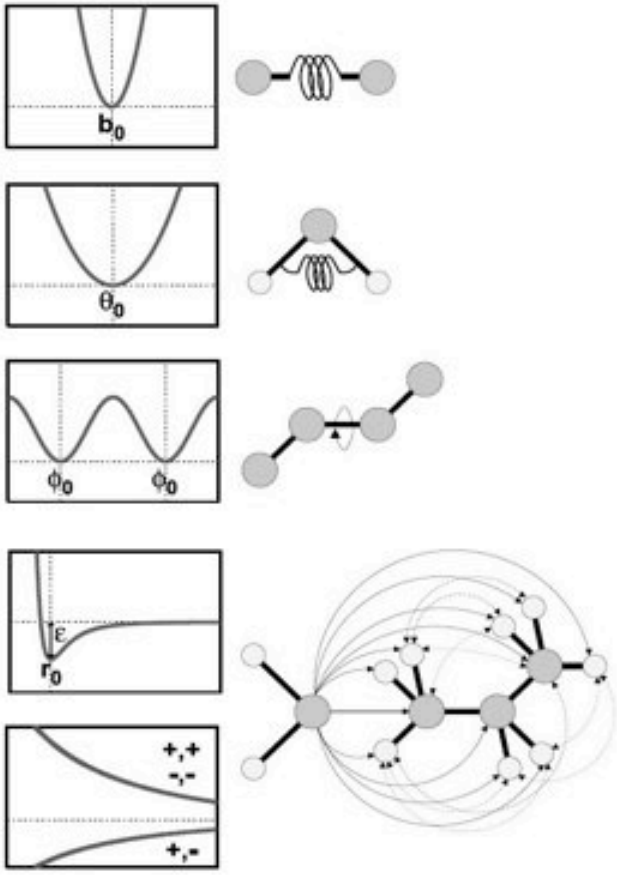
Two main approaches:

(1). **Physics-Based**

(2). **Knowledge-Based**

PHYSICS-BASED POTENTIALS
ENERGY TERMS FROM PHYSICAL THEORY

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



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)

CHARMM P.E. function, see: http://www.charmm.org/

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

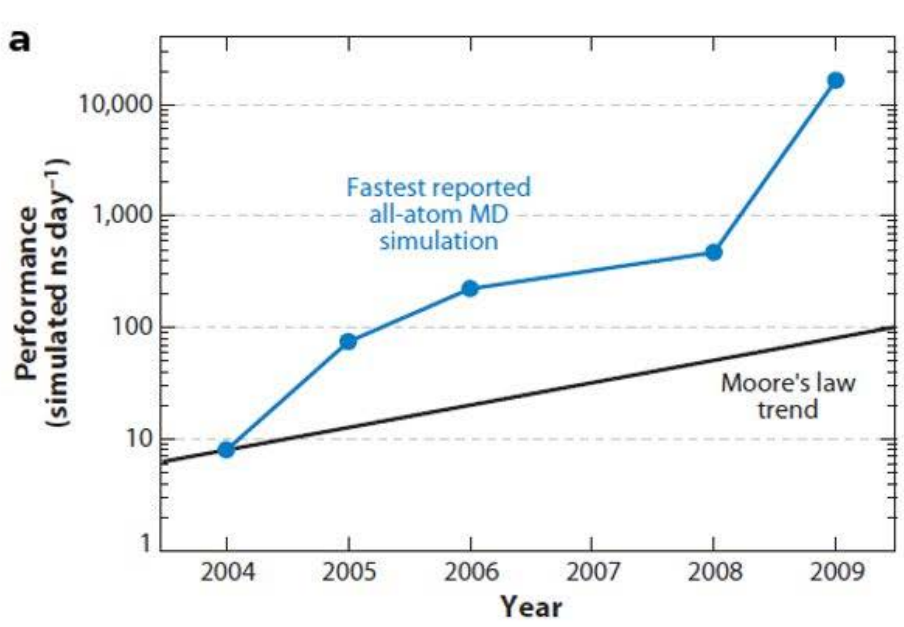
Strengths

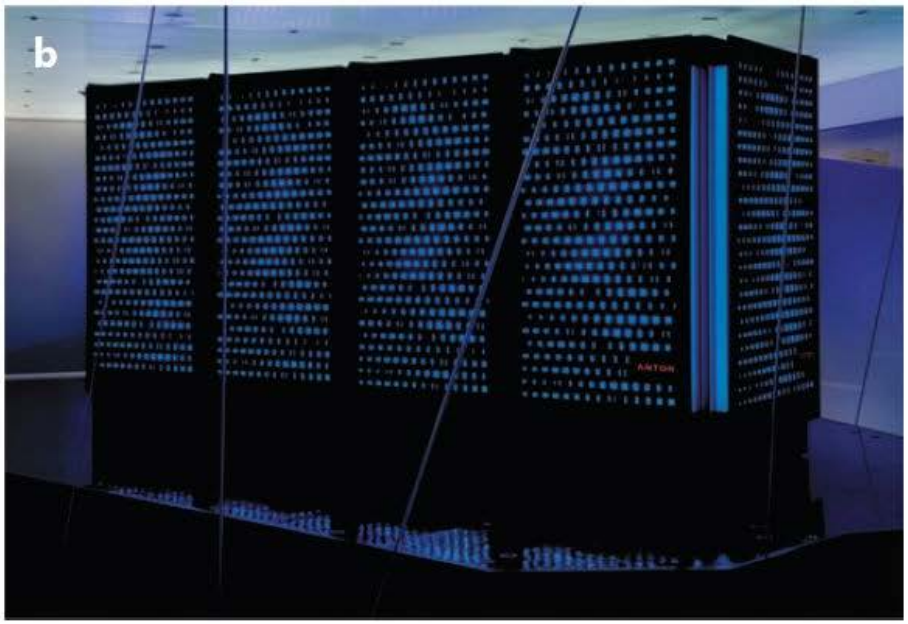
- Interpretable, provides guides to design
- Broadly applicable, in principle at least
- Clear pathways to improving accuracy

Status

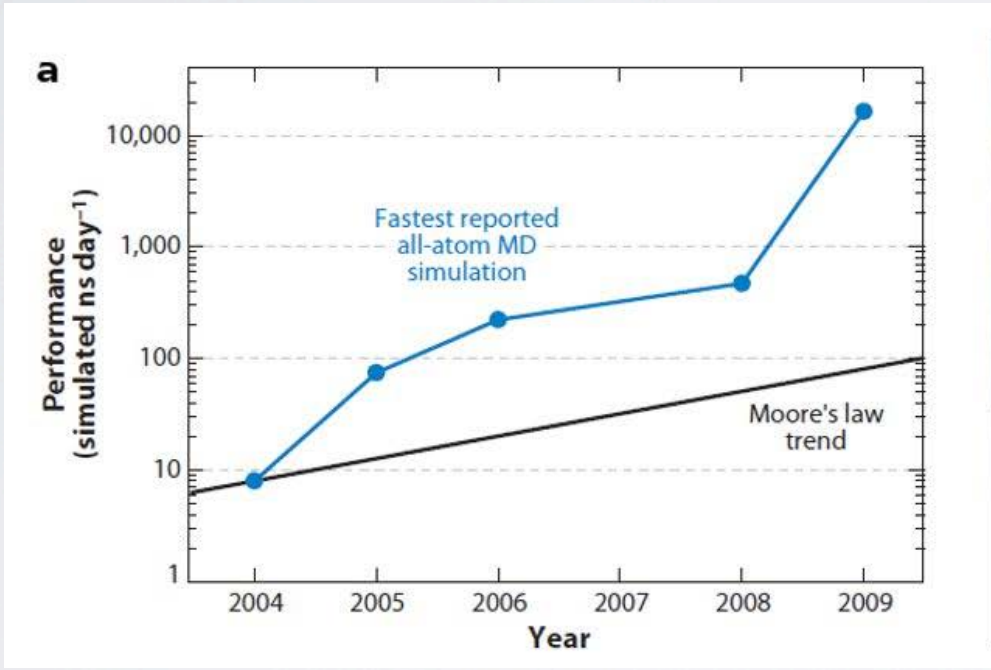
- Useful, far from perfect
- Multiple groups working on fewer, better approxs
  - Force fields, quantum
  - entropy, water effects
- Moore's law: hardware improving

SIDE-NOTE: GPUS AND ANTON SUPERCOMPUTER





SIDE-NOTE: GPUS AND ANTON SUPERCOMPUTER



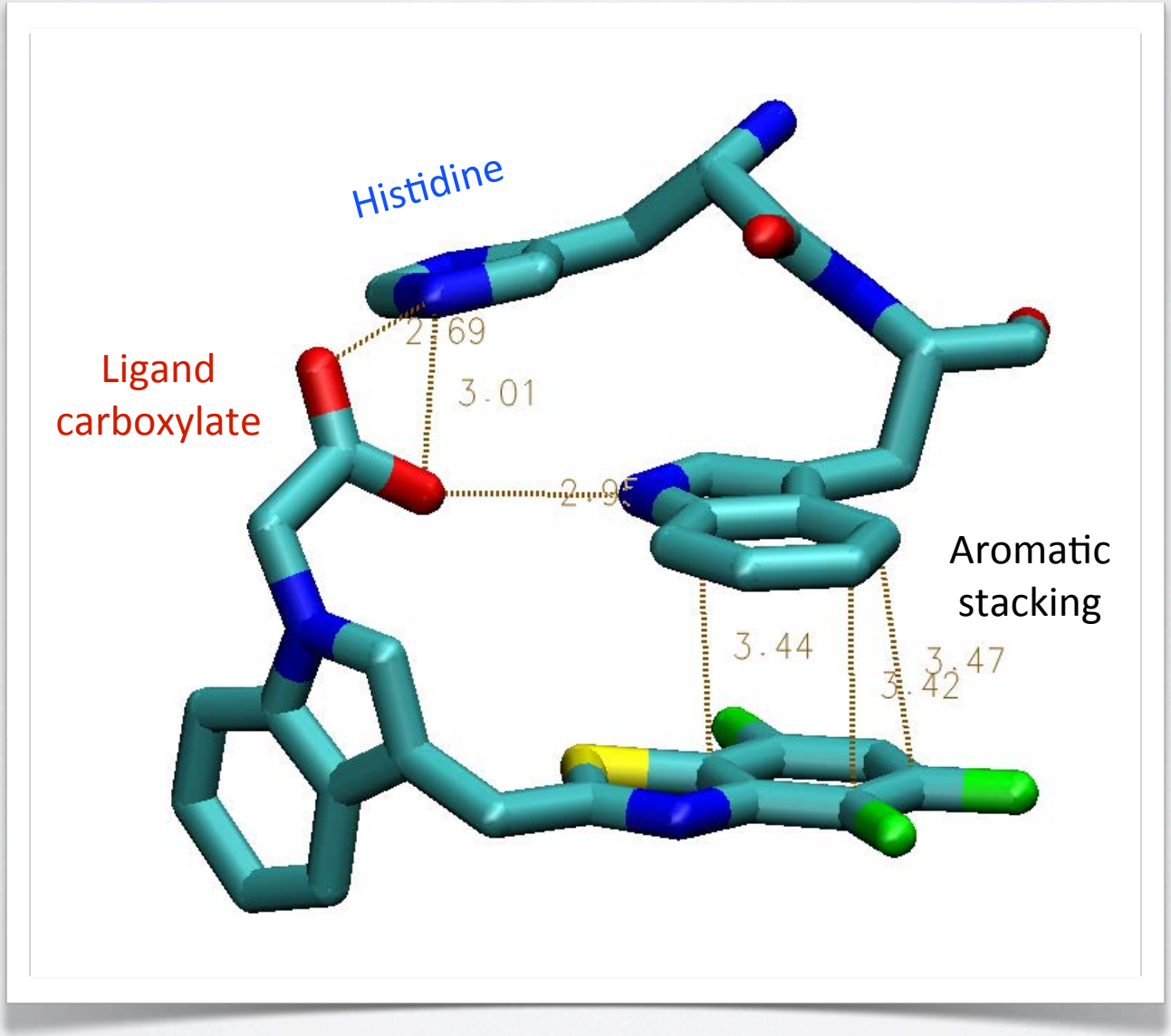
KEY CONCEPT: POTENTIAL FUNCTIONS DESCRIBE A SYSTEMS **ENERGY** AS A FUNCTION OF ITS **STRUCTURE**

Two main approaches:

(1). **Physics-Based**

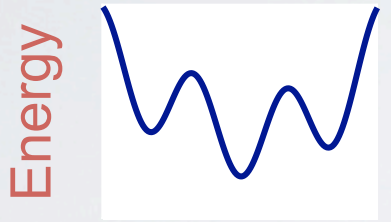
(2). **Knowledge-Based**

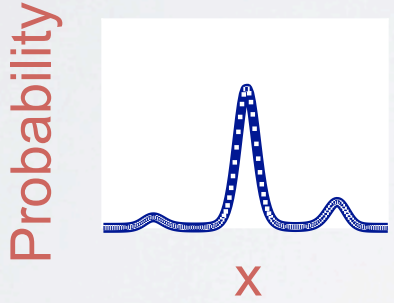
KNOWLEDGE-BASED DOCKING POTENTIALS



ENERGY DETERMINES **PROBABILITY** (STABILITY)

Basic idea: Use probability as a proxy for energy





Boltzmann:

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

Inverse Boltzmann:

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

Example: ligand carboxylate O to protein histidine N

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

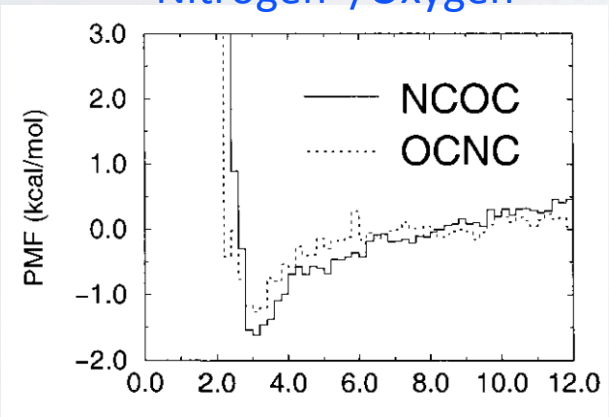
1. For each structure, histogram the distances from O to every histidine N
2. Sum the histograms over all structures to obtain $p(r_{O\text{-}N})$
3. Compute $E(r_{O\text{-}N})$ from $p(r_{O\text{-}N})$

KNOWLEDGE-BASED DOCKING POTENTIALS

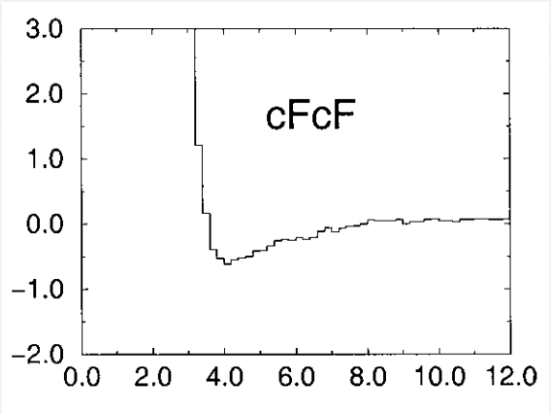
"PMF", Muegge & Martin, J. Med. Chem. (1999) 42:791

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

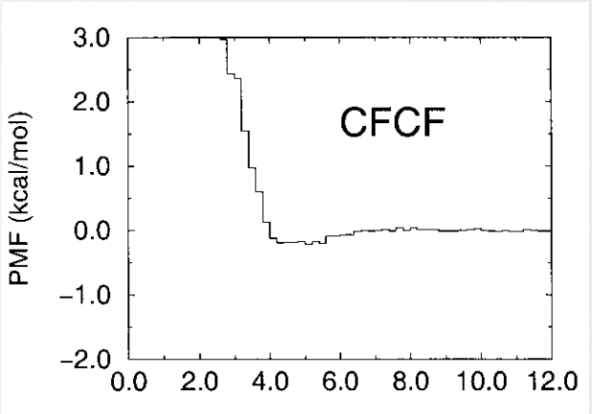
Nitrogen^{+}/Oxygen^{-}



Aromatic carbons



Aliphatic carbons



Atom-atom distance (Angstroms)

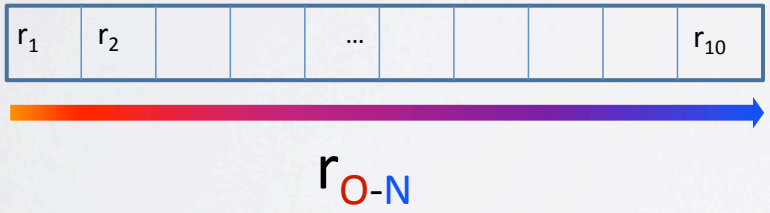
$$E_{prot-lig} = E_{vdw} + \sum_{pairs\ (ij)} E_{type(ij)}(r_{ij})$$

LIMITATIONS OF KNOWLEDGE-BASED POTENTIALS

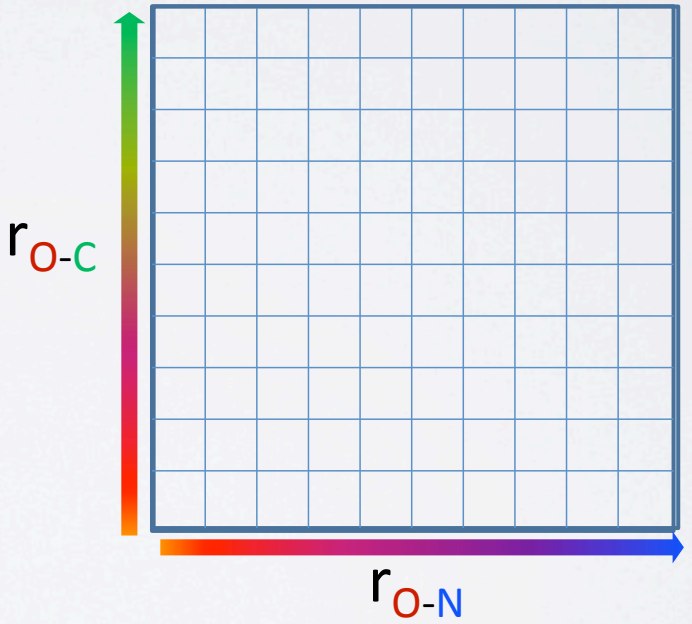
1. Statistical limitations

(e.g., to pairwise potentials)

10 bins for a histogram of O-N distances



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

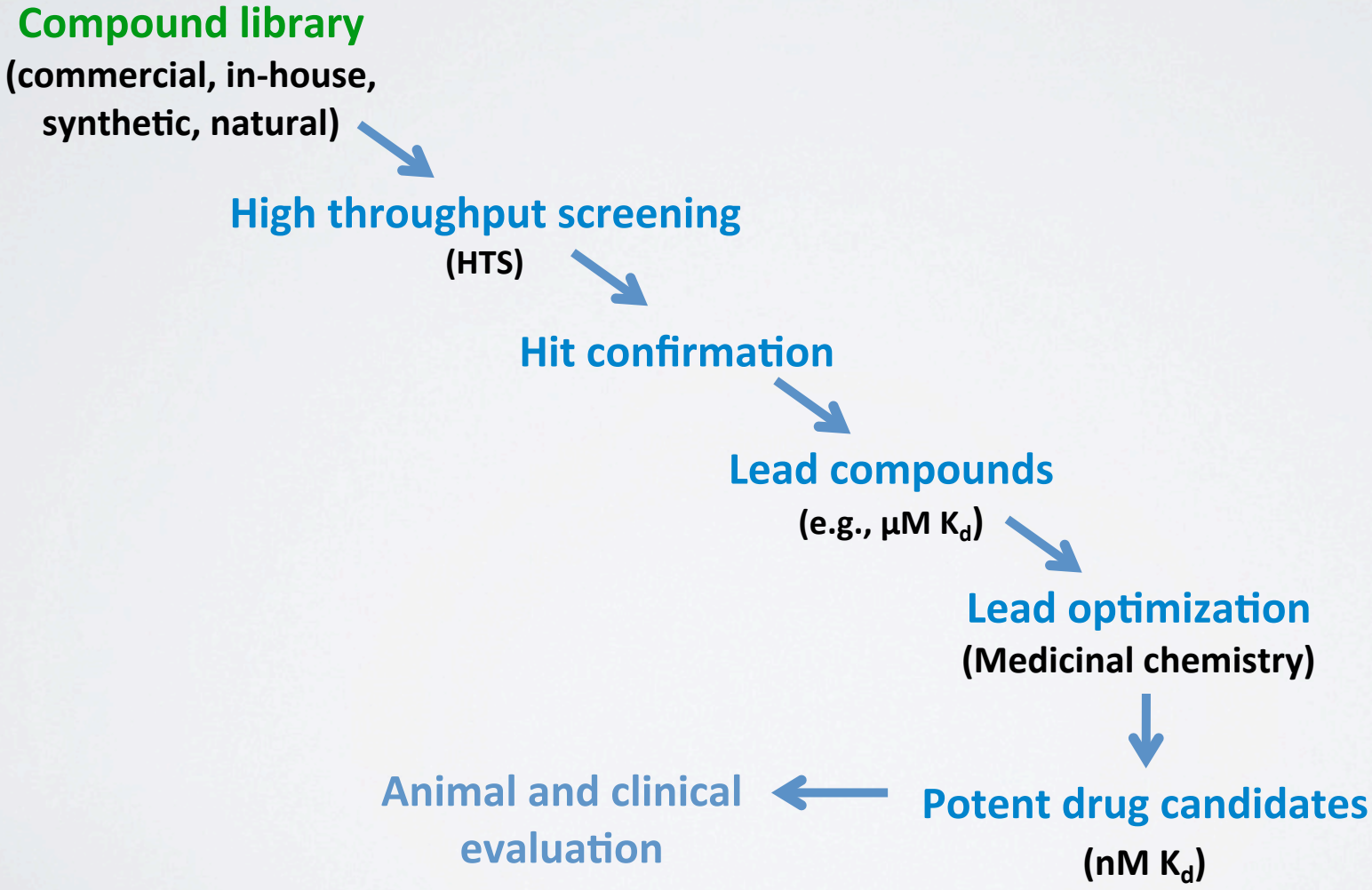
(not always clear how to make improvements)

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



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**

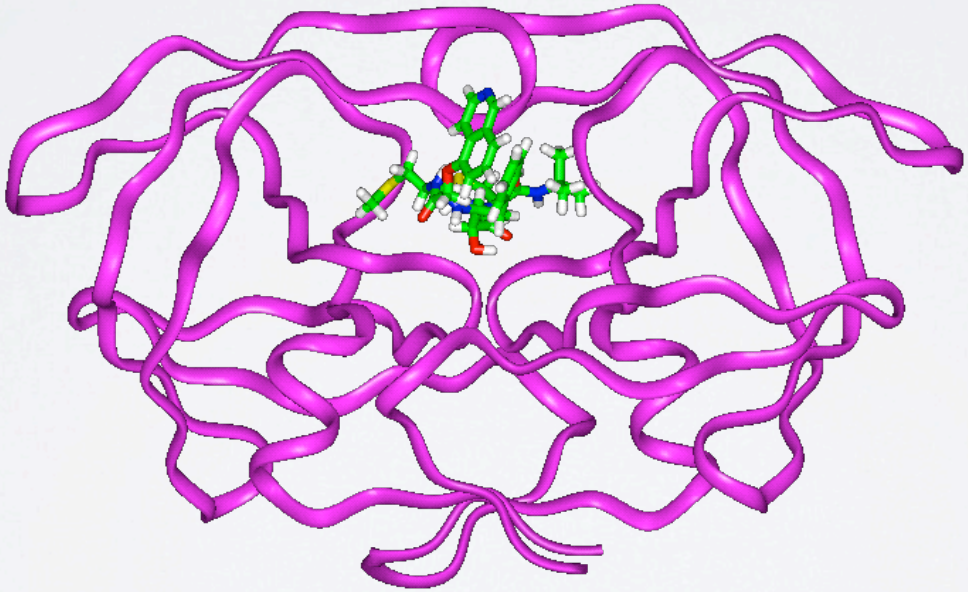
Two main approaches:

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

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

SCENARIO I:
RECEPTOR-BASED DRUG DISCOVERY

Structure of Targeted Protein Known: Structure-Based Drug Discovery



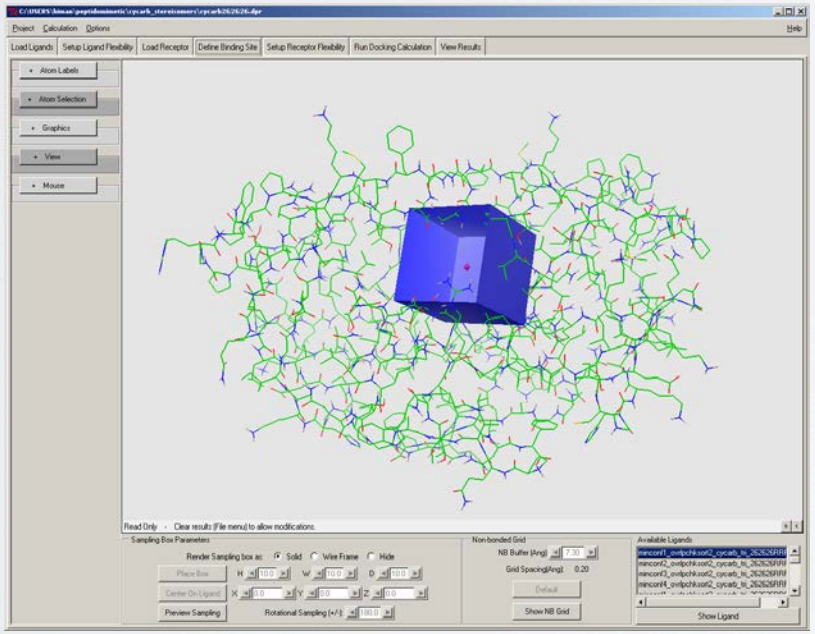
HIV Protease/KNI-272 complex

PROTEIN-LIGAND DOCKING

Structure-Based Ligand Design

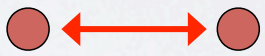
Docking software

Search for structure of lowest energy

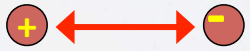


Potential function

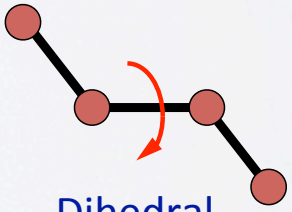
Energy as function of structure



VDW

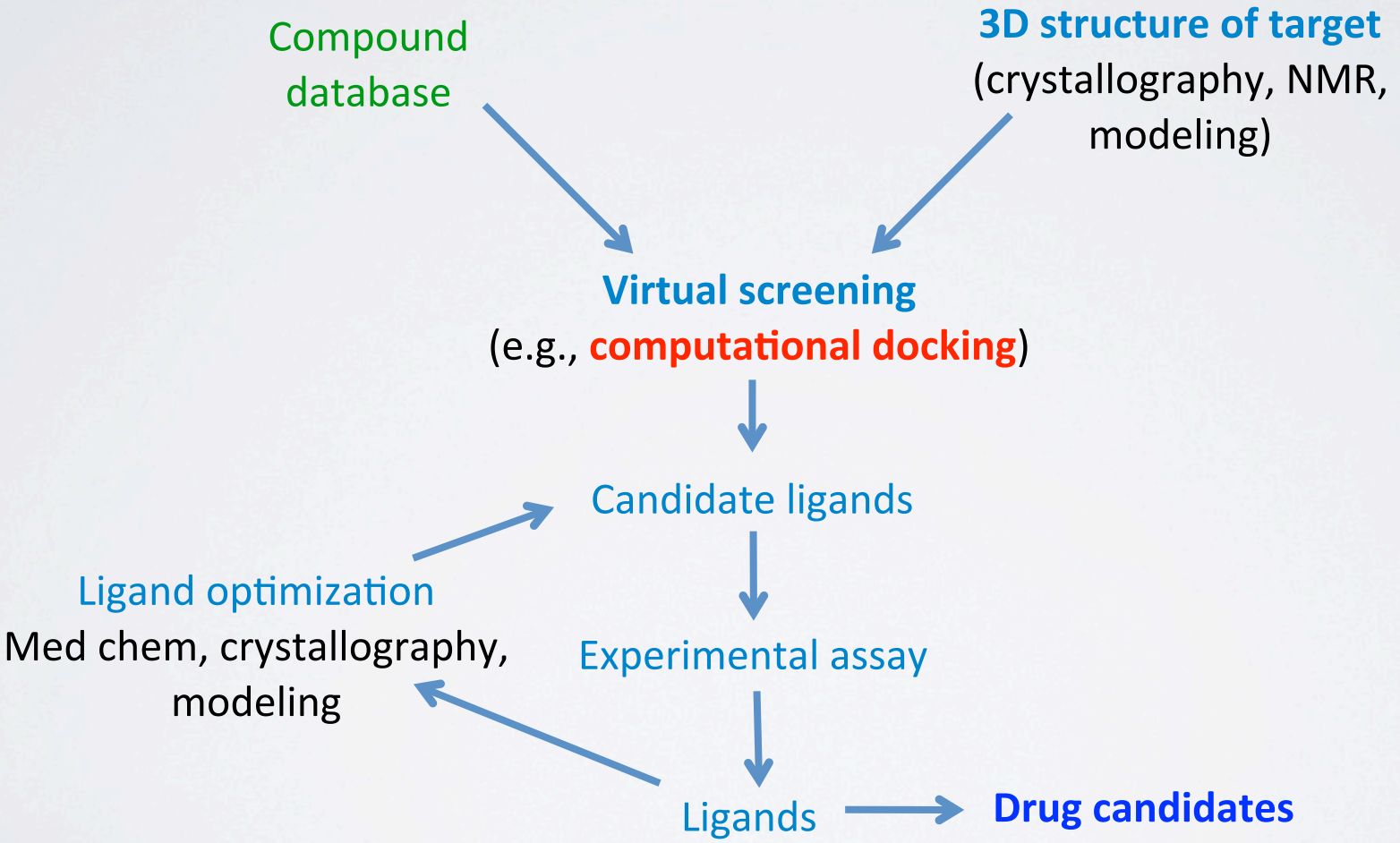


Screened Coulombic

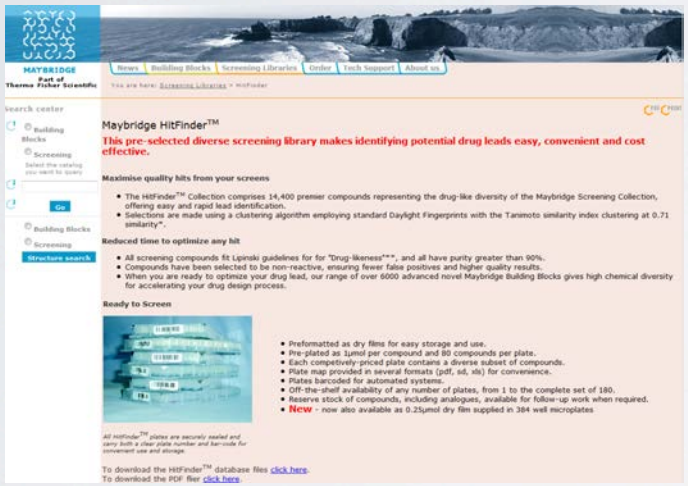


Dihedral

STRUCTURE-BASED VIRTUAL SCREENING

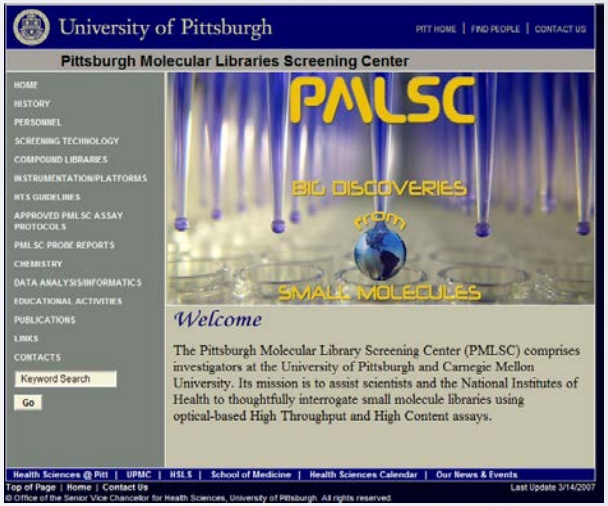


COMPOUND LIBRARIES



Commercial
(in-house pharma)

Government (NIH)



Academia

FRAGMENTAL STRUCTURE-BASED SCREENING

"Fragment" library → **Fragment docking** ← **3D structure of target**

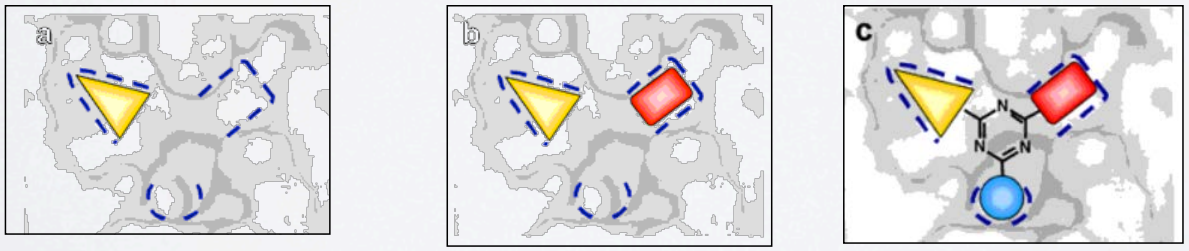
↓

Compound design

↓

Experimental assay and ligand optimization → **Drug candidates**

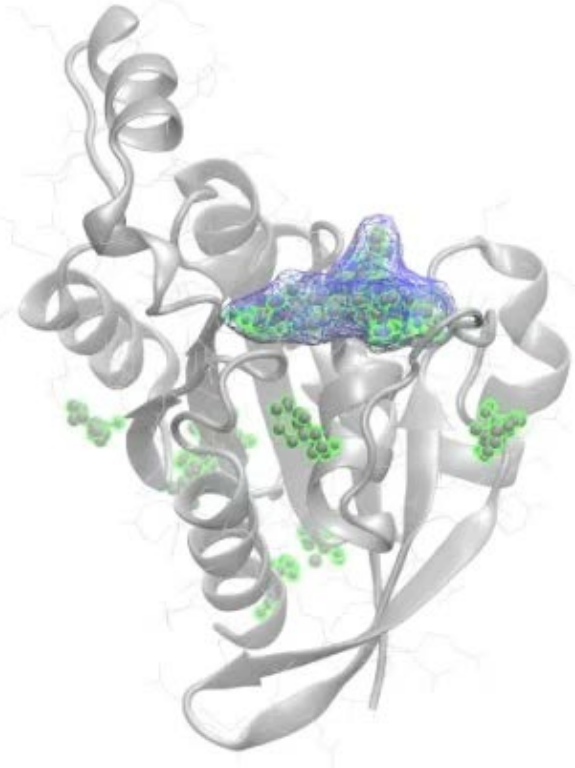
Med chem, crystallography, modeling

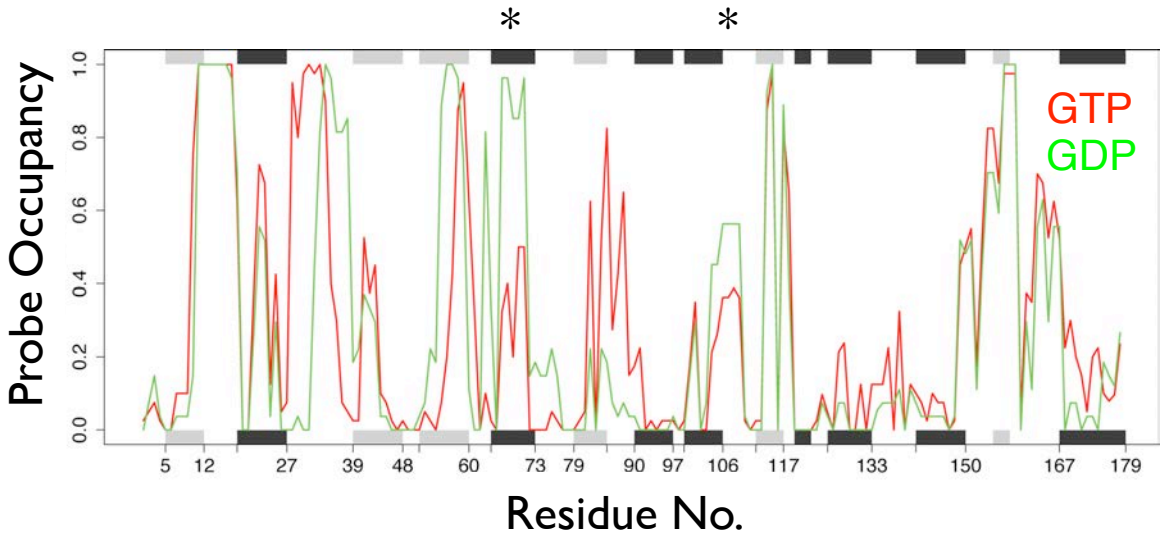


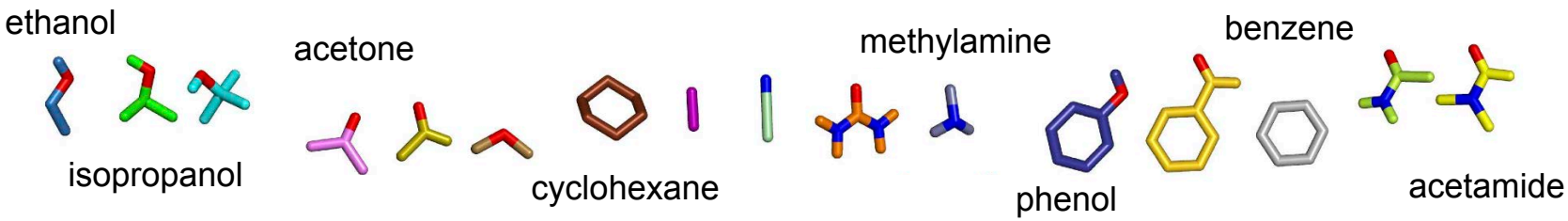
http://www.beilstein-institut.de/bozen2002/proceedings/Jhoti/jhoti.html

Multiple non active-site pockets identified

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



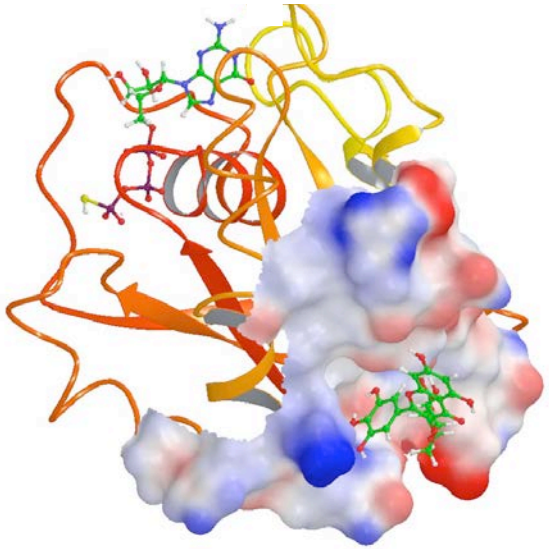


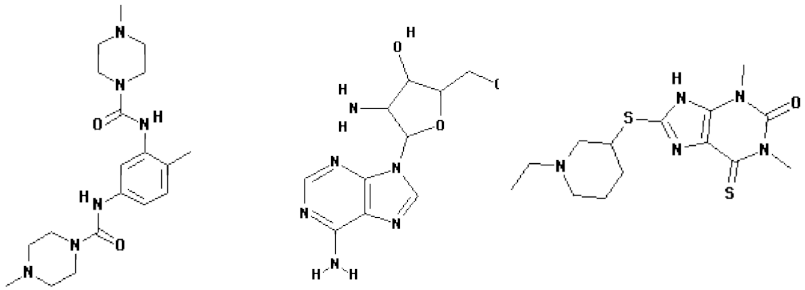


Ensemble docking & candidate inhibitor testing

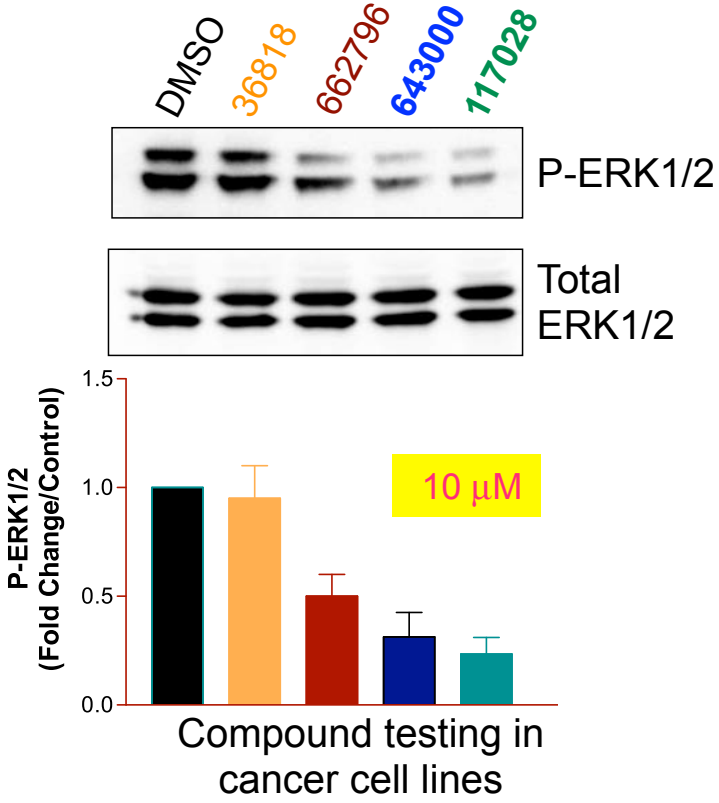
Top hits from ensemble docking against distal pockets were tested for inhibitory effects on basal ERK activity in glioblastoma cell lines.

Ensemble computational docking



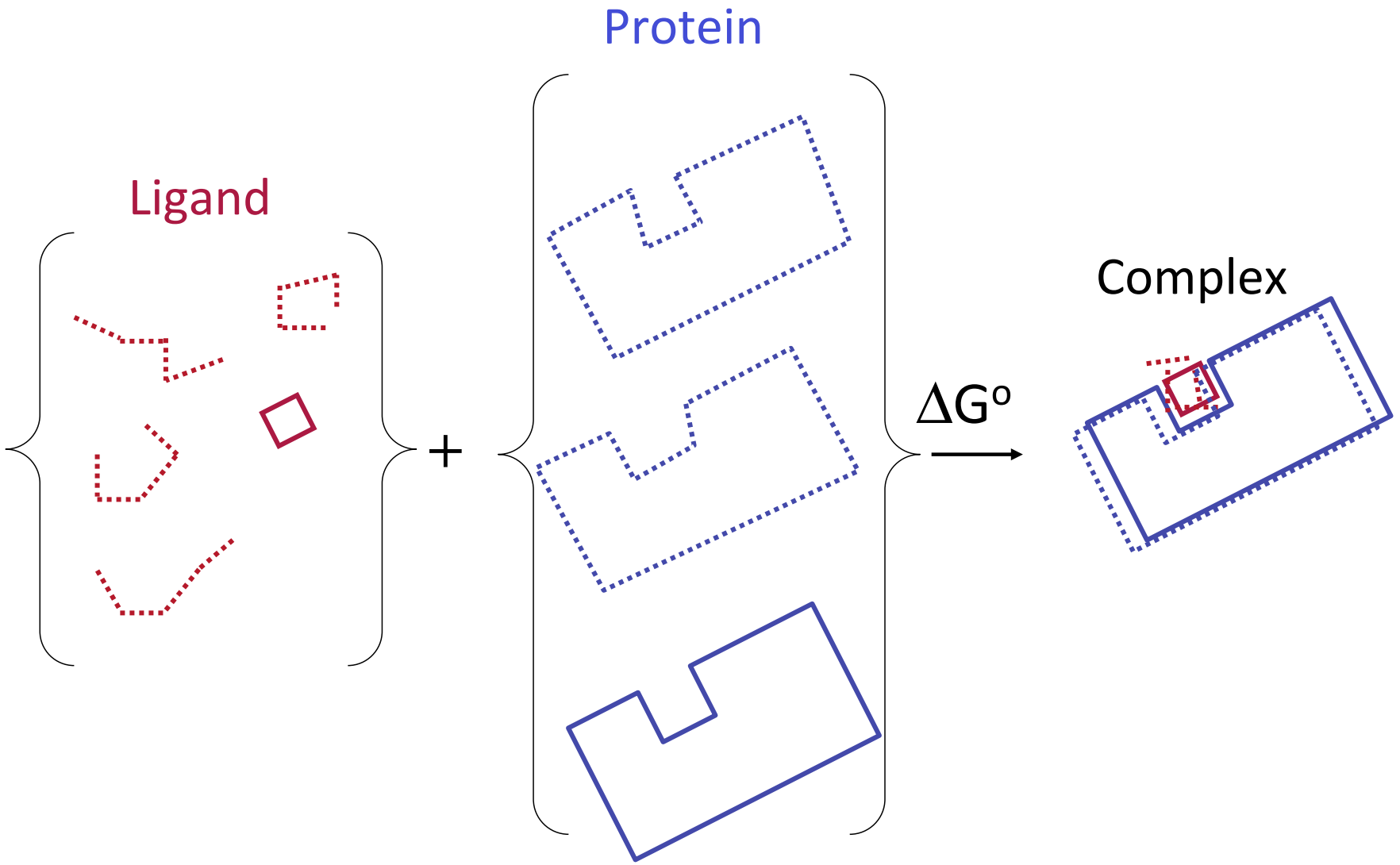


Compound effect on U251 cell line



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:

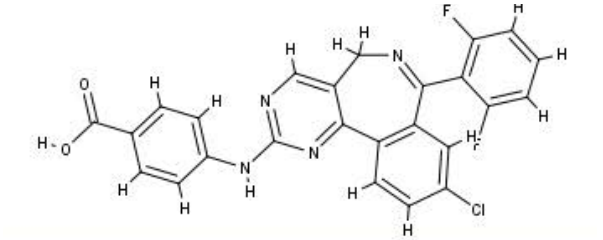
(1). **Receptor/Target-Based**

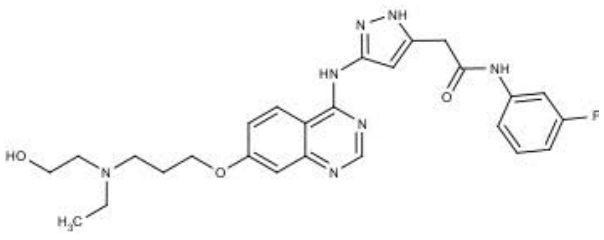
(2). **Ligand/Drug-Based**

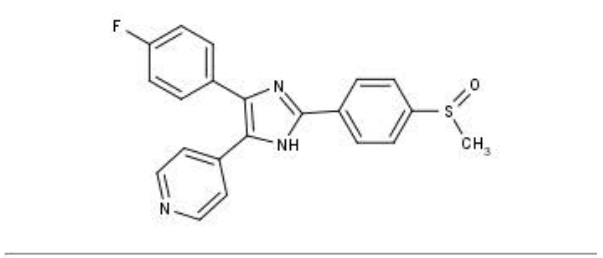
Scenario 2

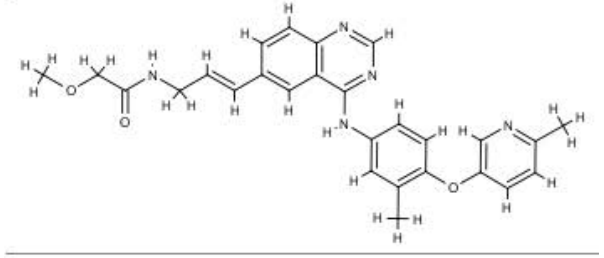
Structure of Targeted Protein Unknown: Ligand-Based Drug Discovery

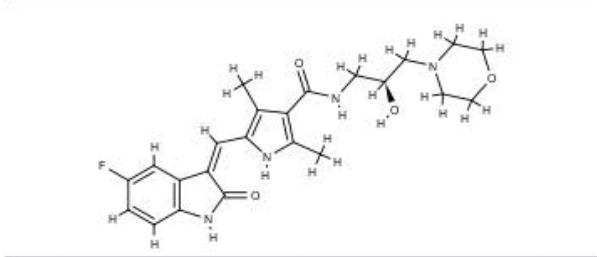
e.g. MAP Kinase Inhibitors

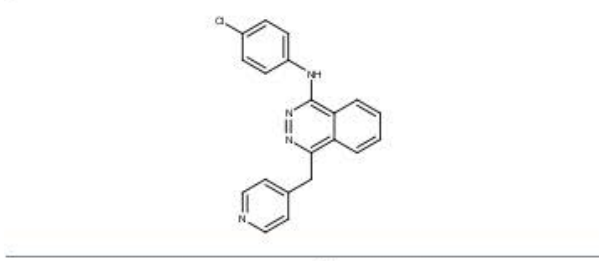


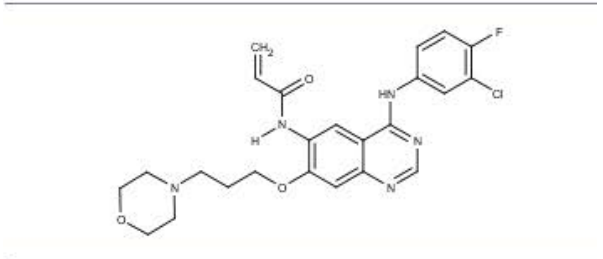


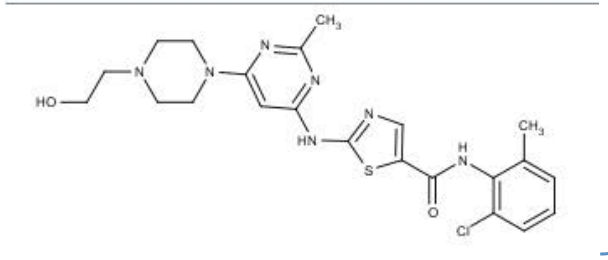












Using knowledge of existing inhibitors to discover more

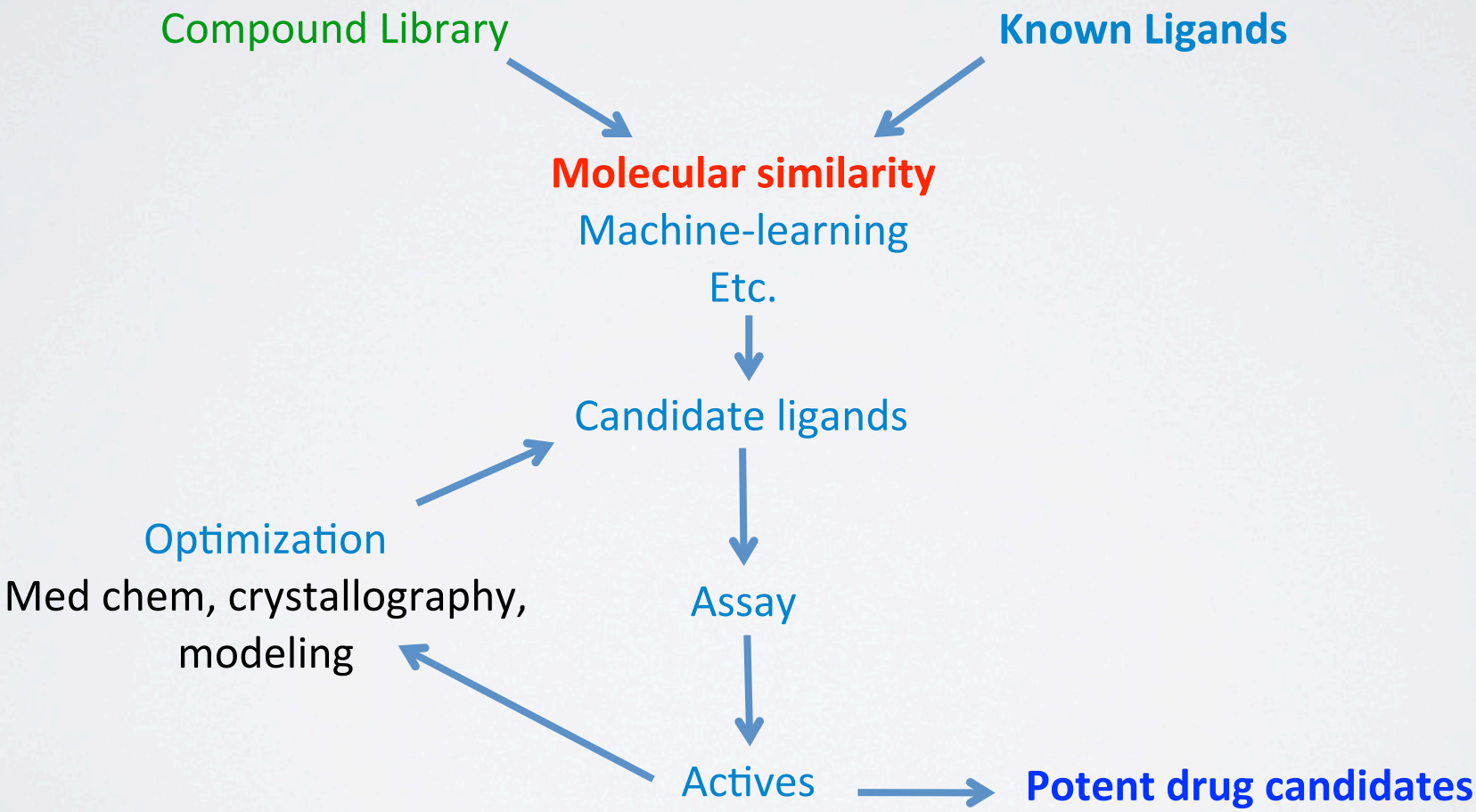
Why Look for Another Ligand if You Already Have Some?

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

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

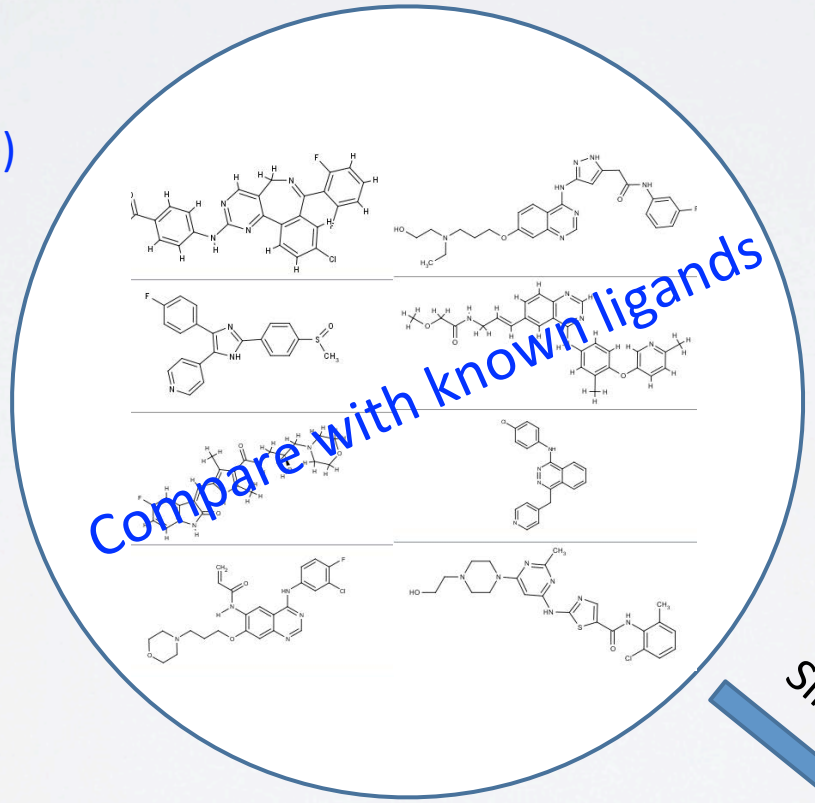
An high-affinity ligand is toxic, is not well-absorbed, etc.

LIGAND-BASED VIRTUAL SCREENING



CHEMICAL SIMILARITY
LIGAND-BASED DRUG-DISCOVERY

Compounds
(available/synthesizable)



Different

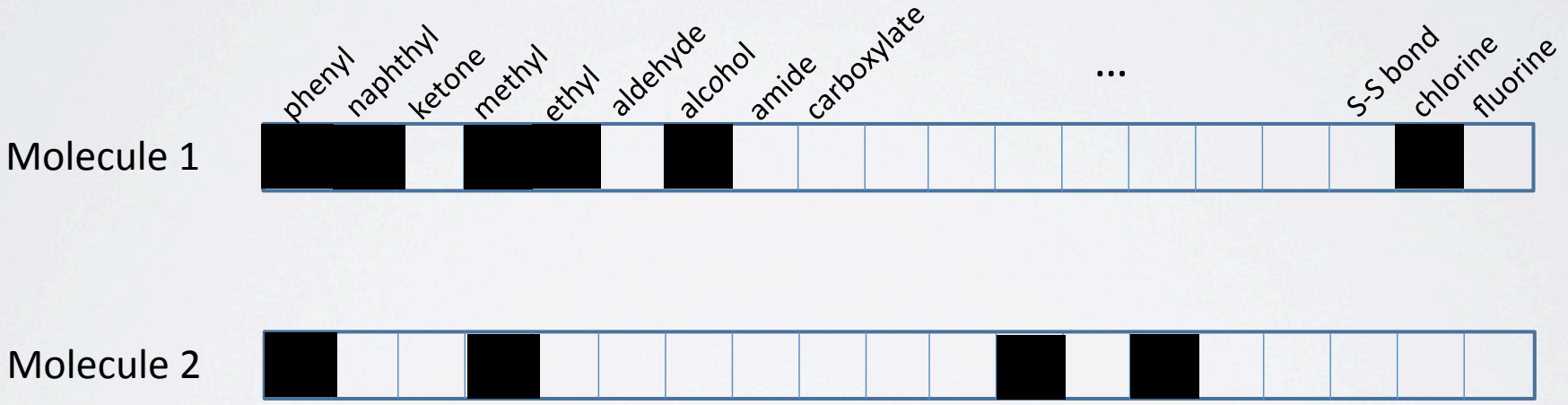
Don't bother

Similar

Test experimentally

CHEMICAL FINGERPRINTS

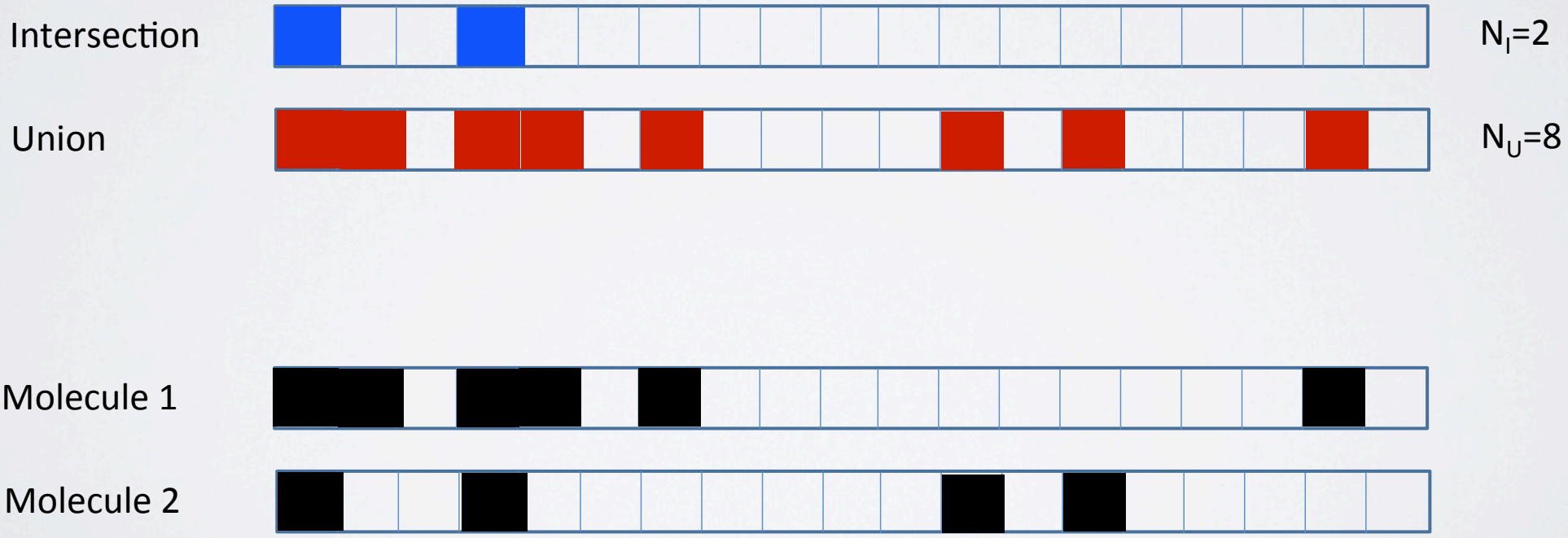
BINARY STRUCTURE KEYS



CHEMICAL SIMILARITY FROM FINGERPRINTS

Tanimoto Similarity or Jaccard Index, T

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



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

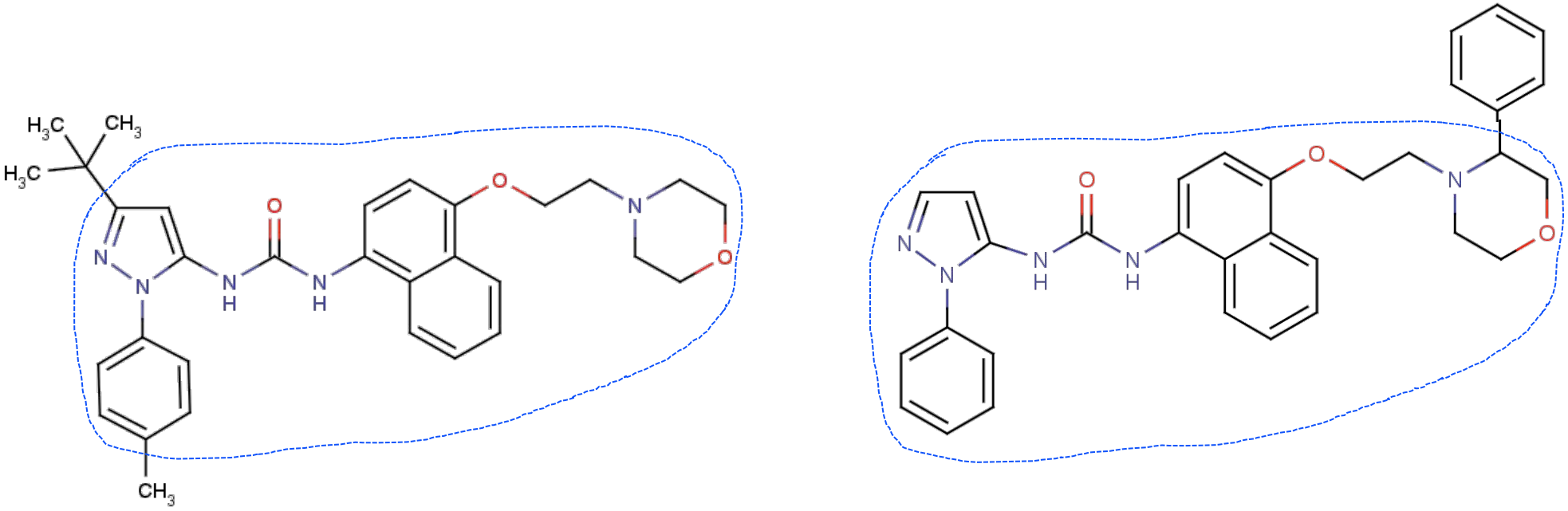
Ligand shape and common substructures

Pharmacophore models

Chemical descriptors

Statistics and machine learning

Maximum Common Substructure

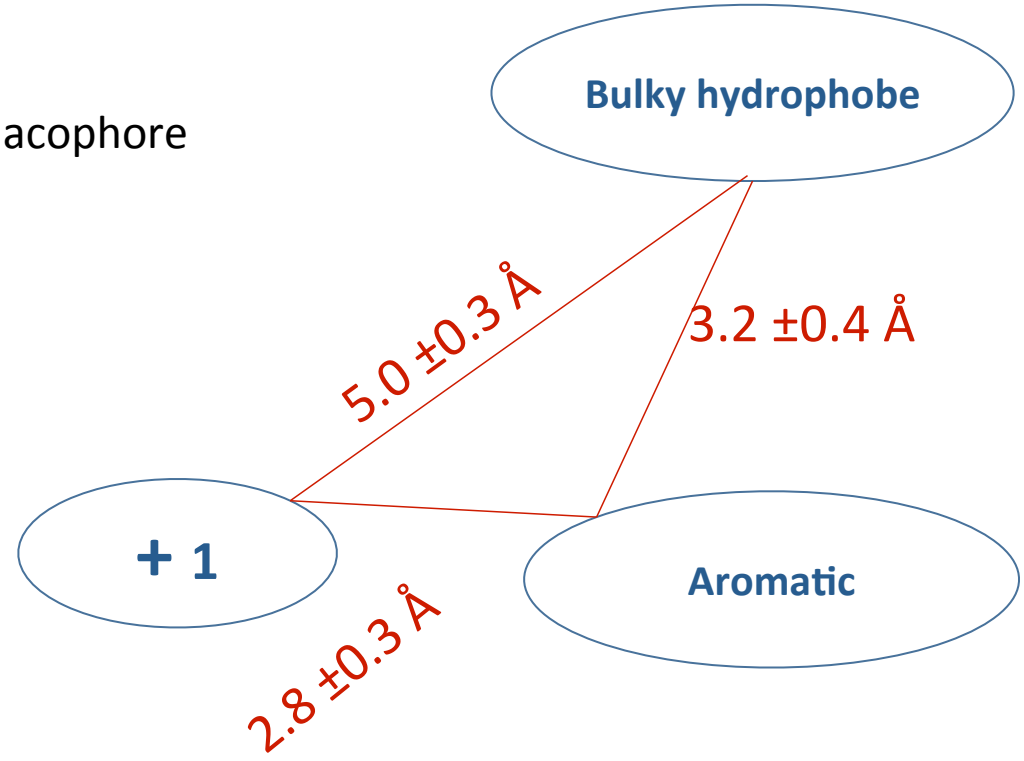


$N_{common}=34$

Pharmacophore Models

Φάρμακο (drug) + Φορά (carry)

A 3-point pharmacophore

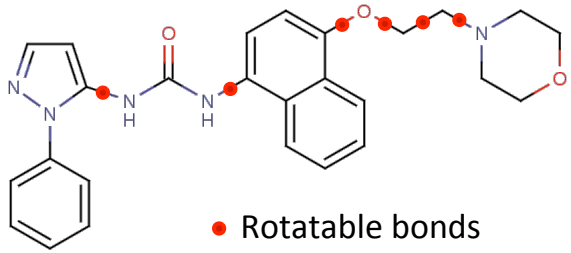


Molecular Descriptors

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 clogP)



Topological

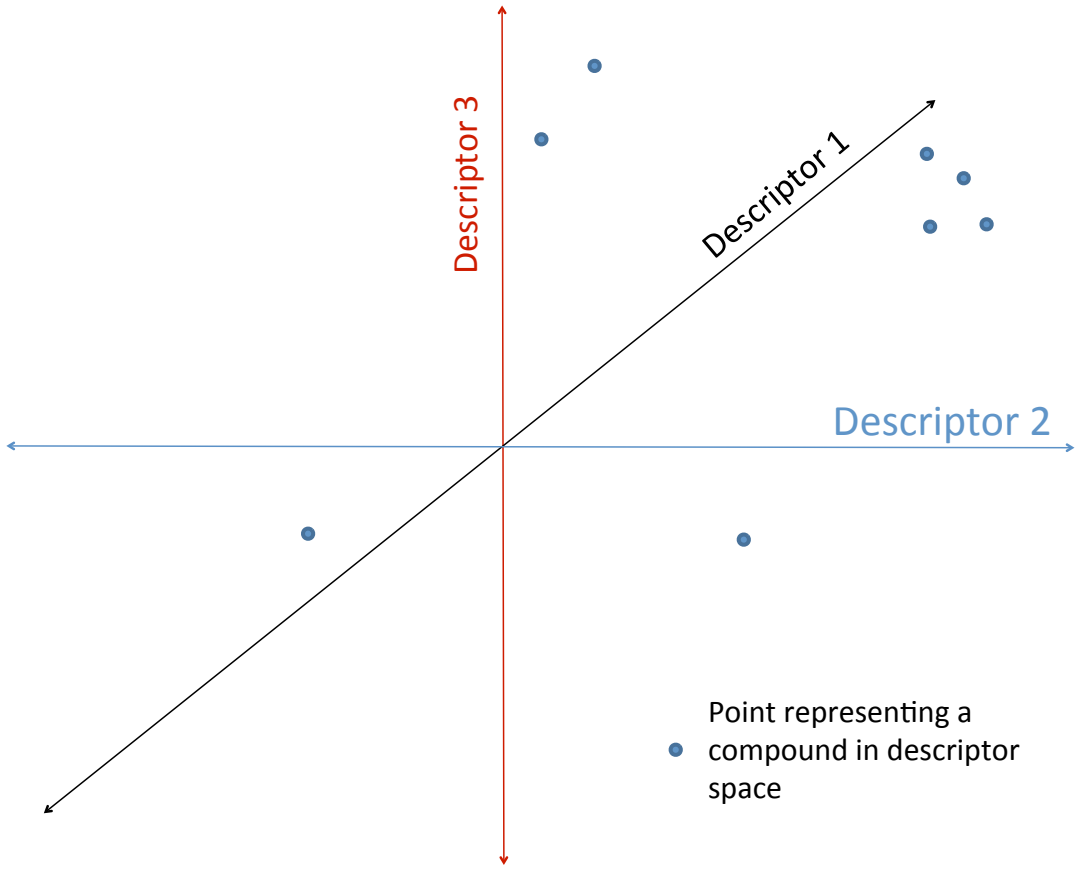
- branching index
- measures of linearity vs interconnectedness

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)

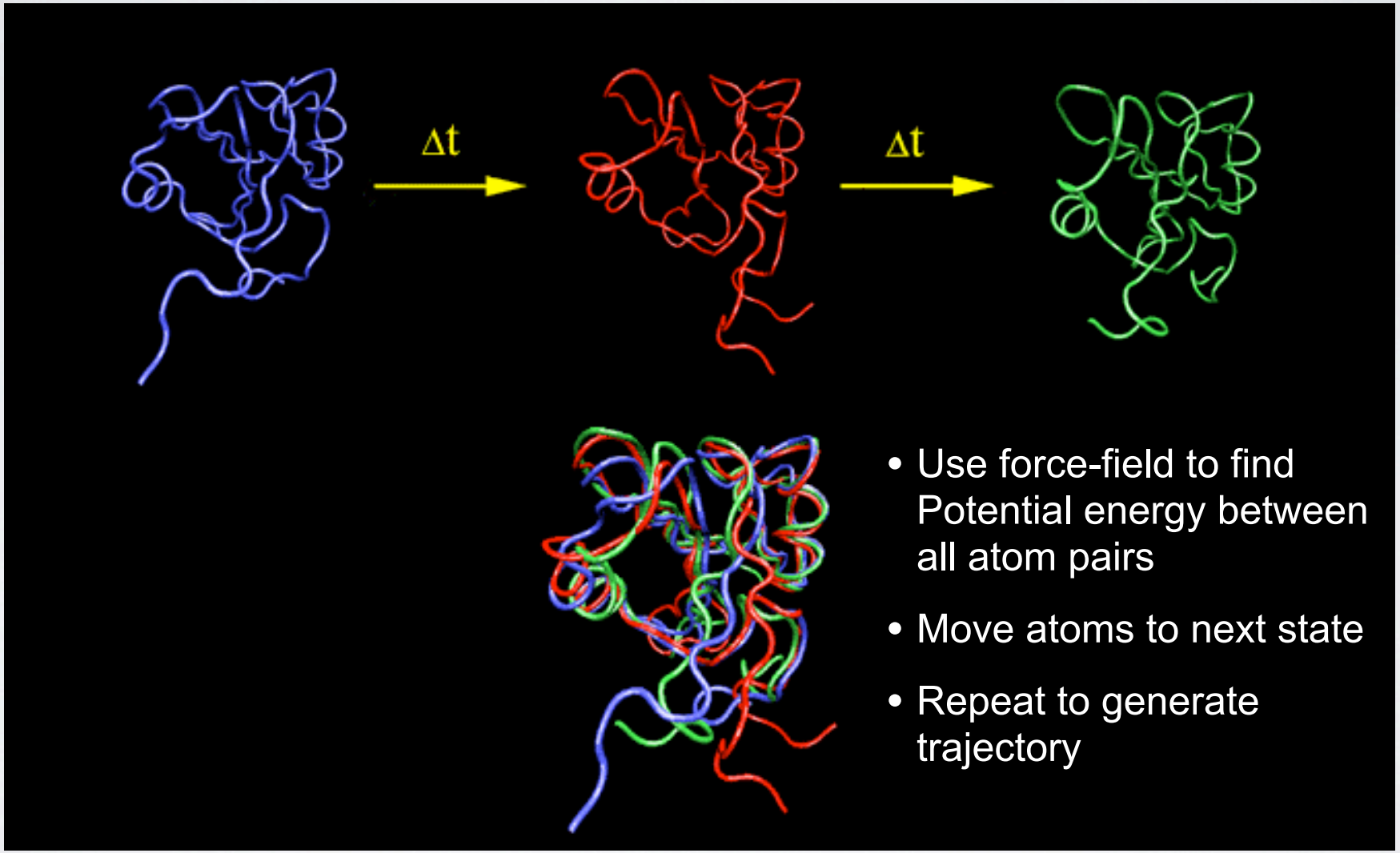
Ligand-protein databases, machine-readable chemical formats

Ligand similarity and beyond

PREDICTING FUNCTIONAL DYNAMICS MOLECULAR DYNAMICS SIMULATIONS

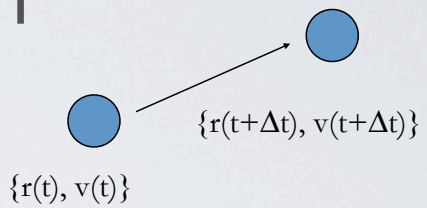
- Proteins are intrinsically flexible molecules with internal motions that are often intimately coupled to their biochemical function.
  - E.g. ligand and substrate binding, allosteric regulation
- Thus knowledge of dynamics can provide a deeper understanding of the mapping of structure to function.
- 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

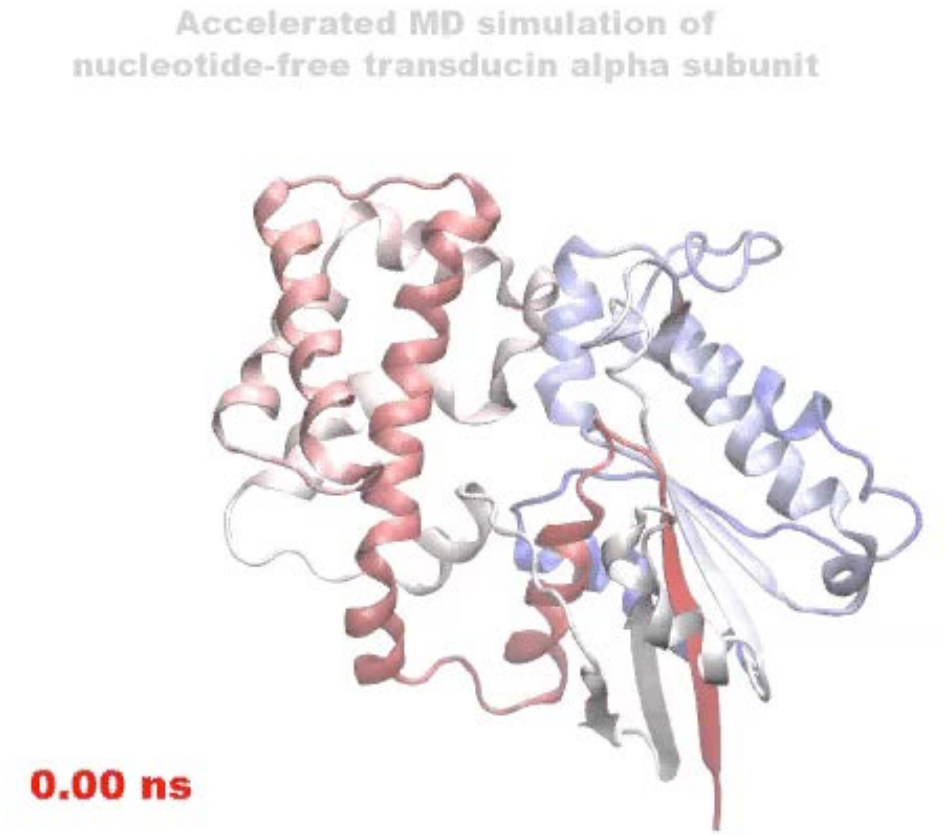
Leapfrog algorithm

1 solve for a_i at t using: $-\frac{dE}{dr_i} = F_i = m_i\, a_i(t)$

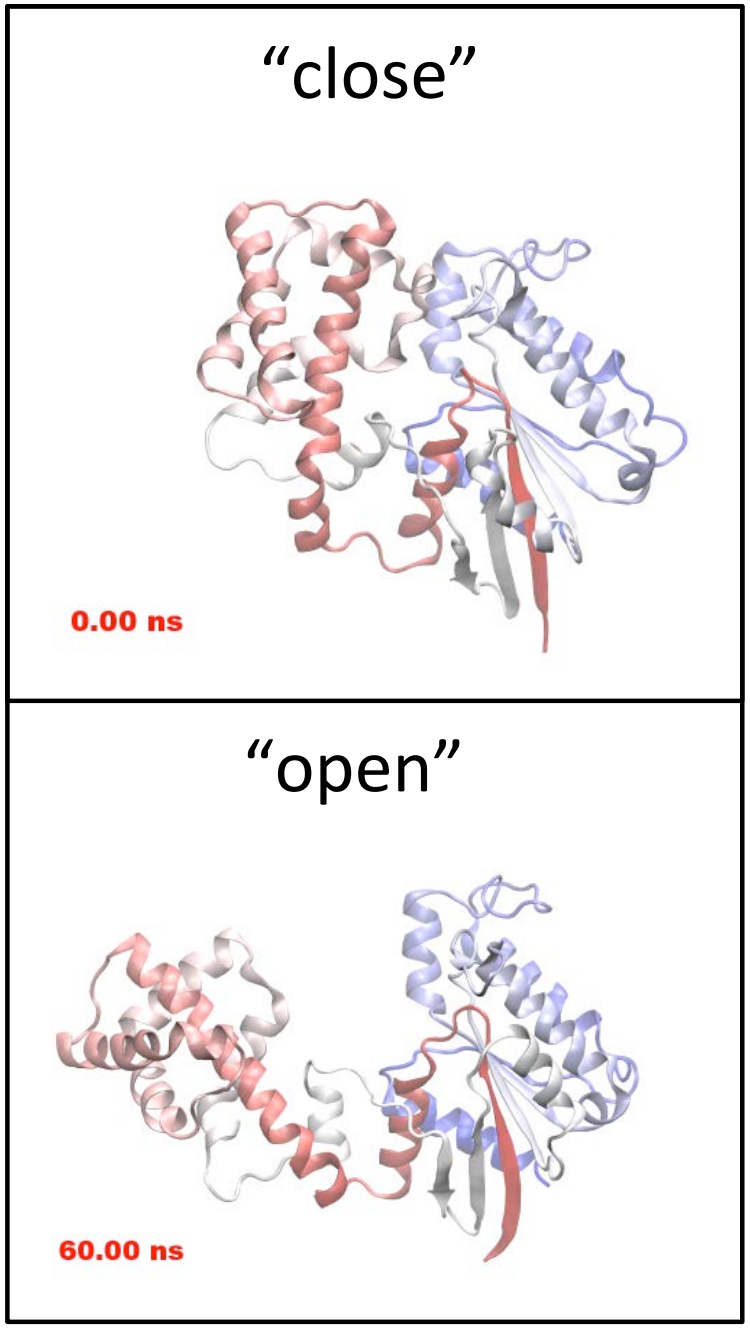
2 update v_i at $t + \Delta t/2$ using: $v_i(t + \Delta t/2) = v_i(t - \Delta t/2) + a_i(t)\,\Delta t$

3 update r_i at $t + \Delta t$ using: $r_i(t + \Delta t) = r_i(t) + v_i(t + \Delta t/2)\,\Delta t$

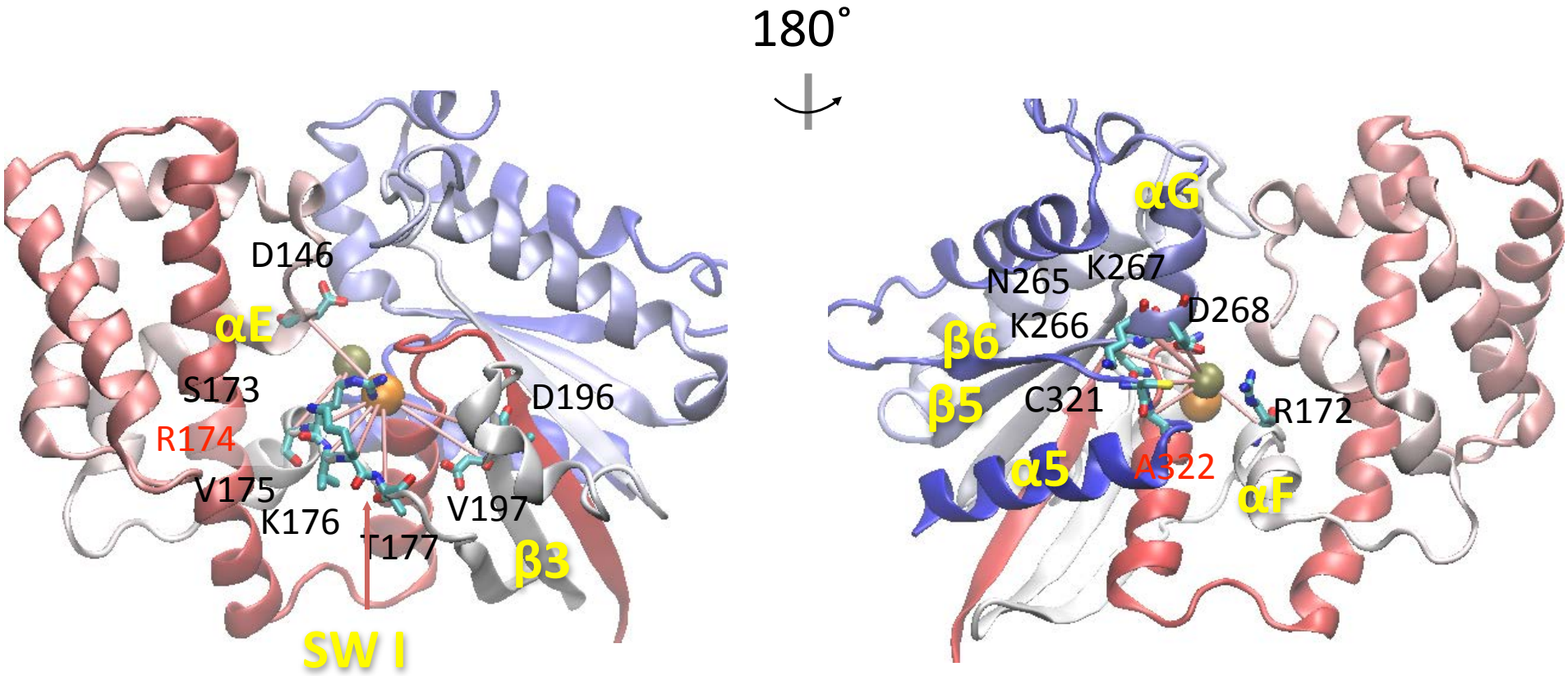
MD Prediction of Functional Motions



Yao and Grant, Biophys J. (2013)



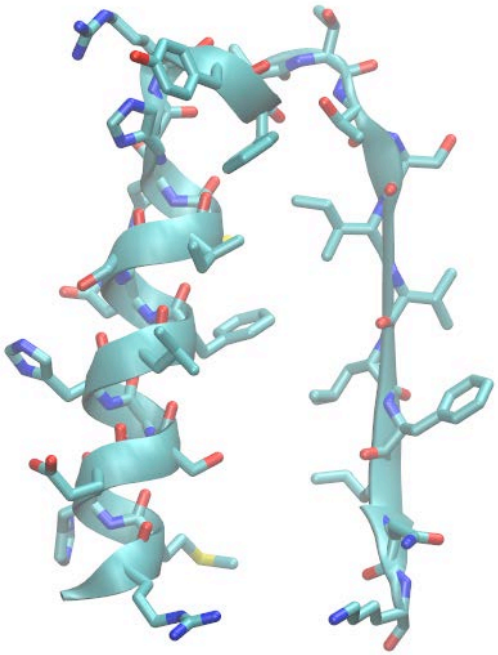
Key Residues Mediating Coupling Between Residues And Nucleotide



Yao and Grant, Biophys J. (2013)

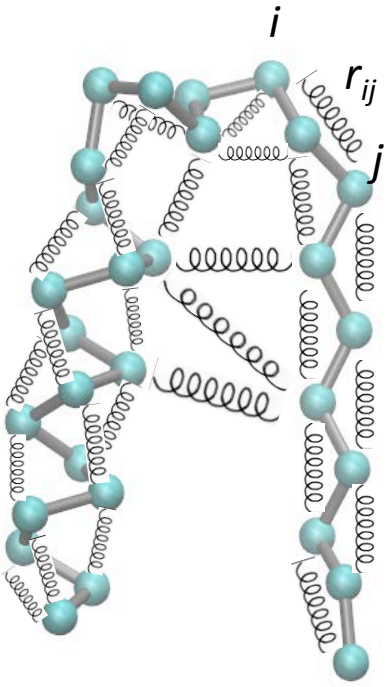
Normal Mode Analysis (NMA)

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



Atomic

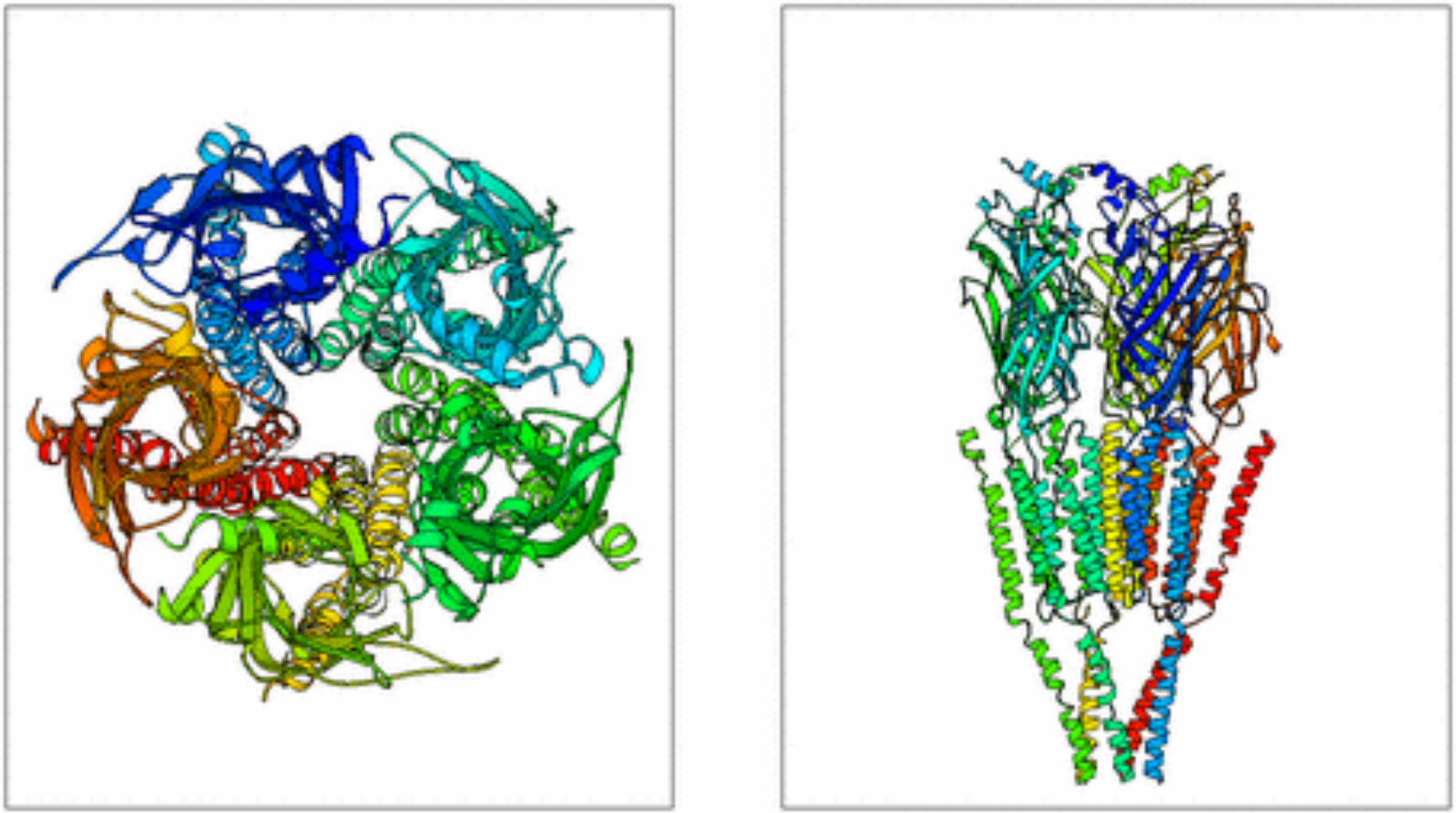
C. G.



a. a.

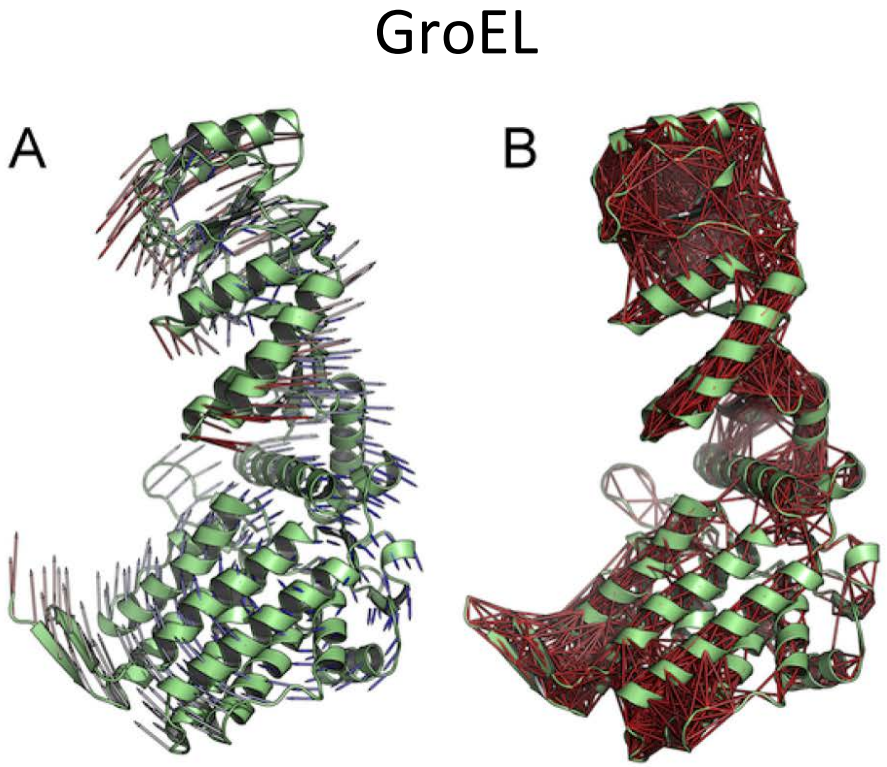
- 1 bead / 1 amino acid
- Connected by springs

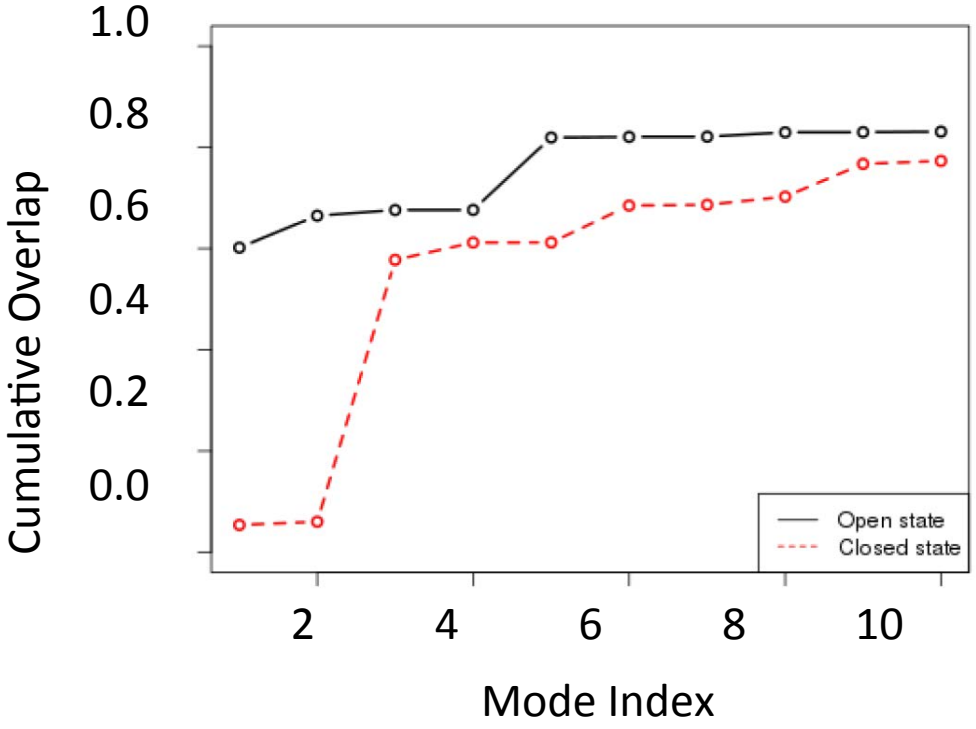
Normal mode of acetylcholine receptor



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

Problems in Conventional ENM-NMA

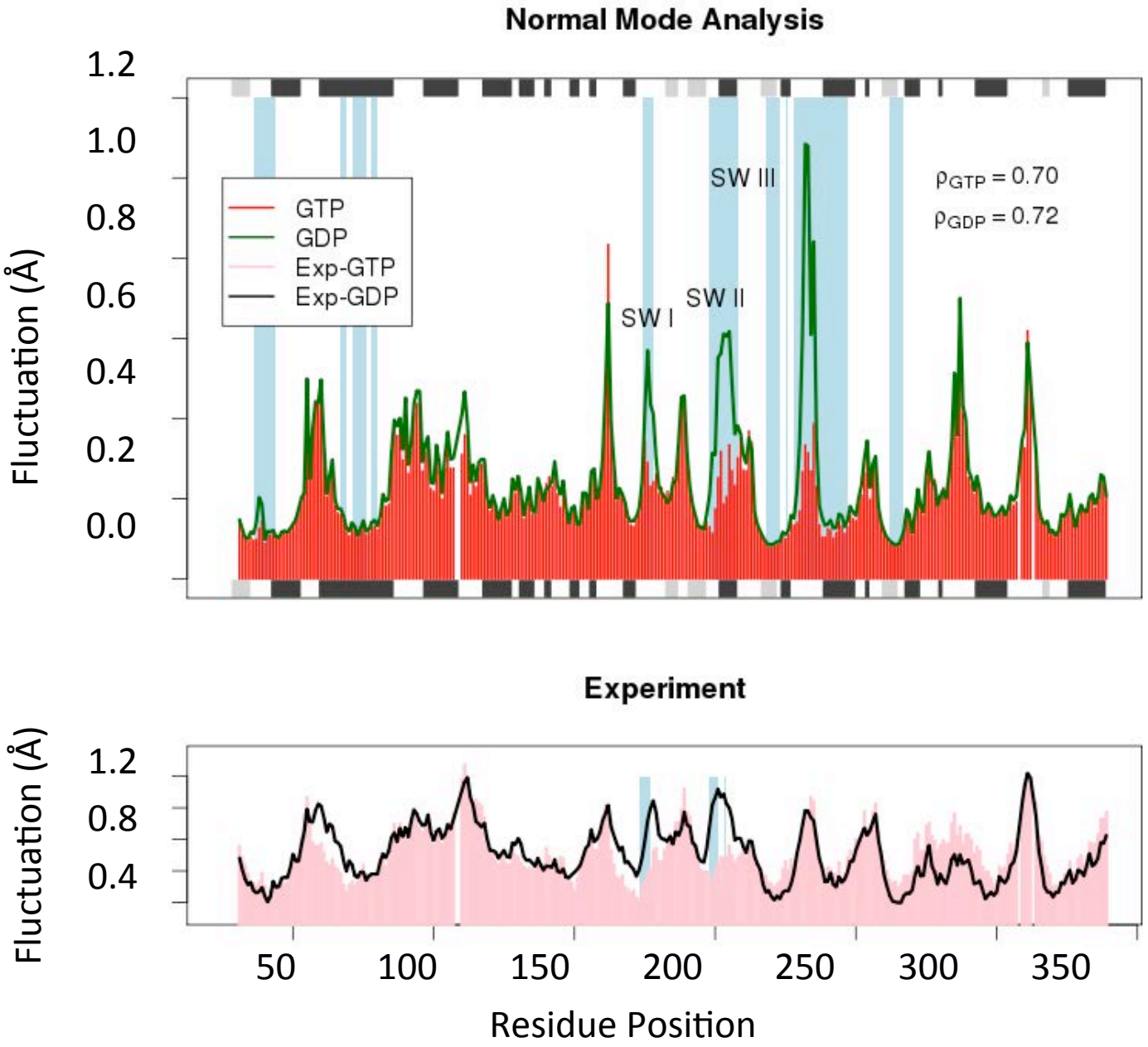




- Work well for elongated multi-domain systems such as GroEL
- But, results are **dependent** on the **input structure** - open forms work best!

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

NMA Predicts High Flexibility in Functional Regions



Lars, Yao & Grant, *in preparation*

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?