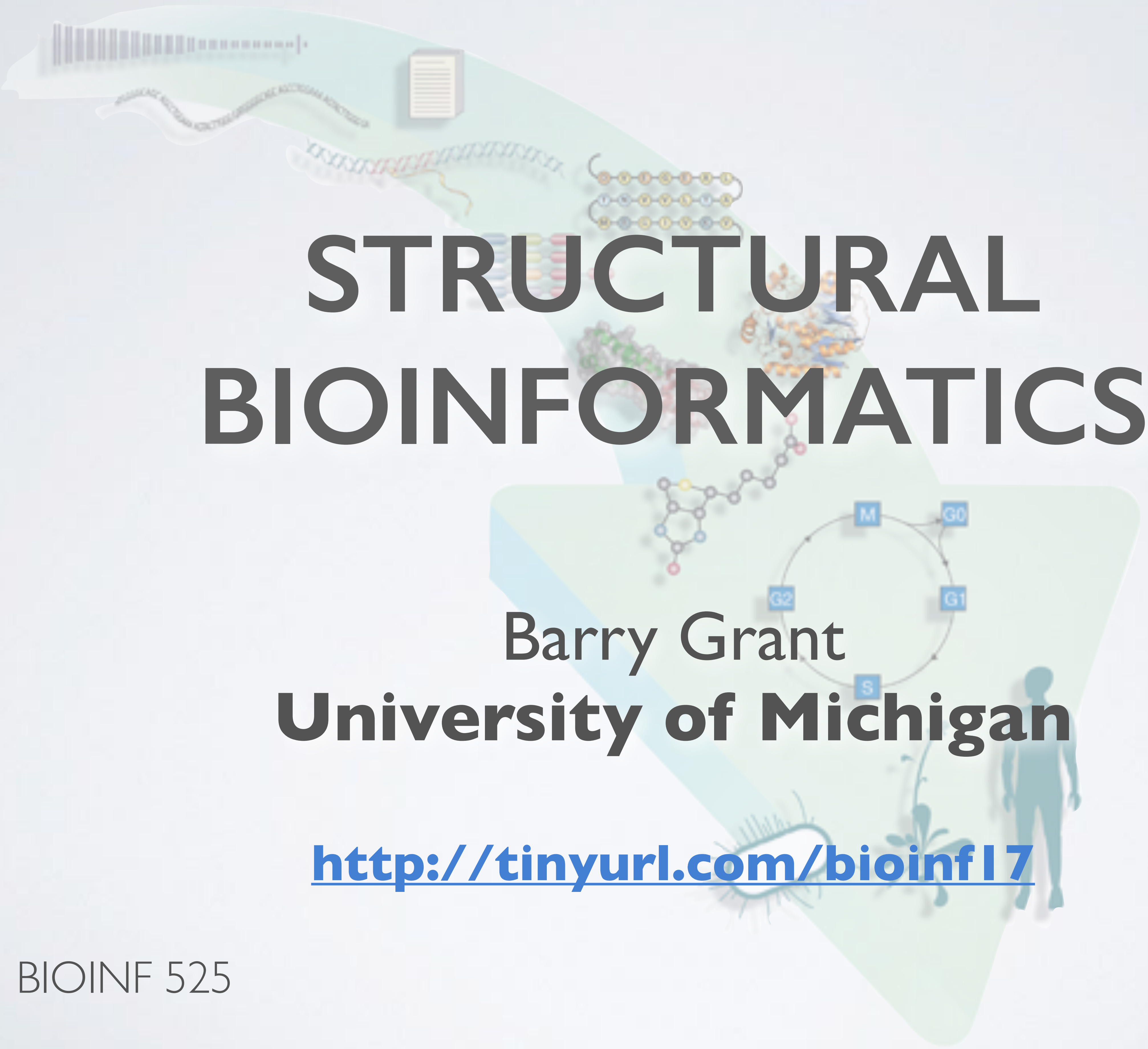


STRUCTURAL BIOINFORMATICS

Barry Grant

University of Michigan

http://tinyurl.com/bioinf17

BIOINF 525

24-Jan-2017

MODULE OVERVIEW

Objective: Provide an introduction to the practice of bioinformatics as well as a practical guide to using common bioinformatics databases and algorithms

1.1. *Introduction to Bioinformatics*

1.2. *Sequence Alignment and Database Searching*

1.3 *Structural Bioinformatics*

1.4 *Genome Informatics: High Throughput Sequencing Applications and Analytical Methods*

WEEK TWO REVIEW

☑ **Answers to last weeks homework:**

Answers week 2

☑ **Muddy Point Assessment** (Only 25 responses):

Responses

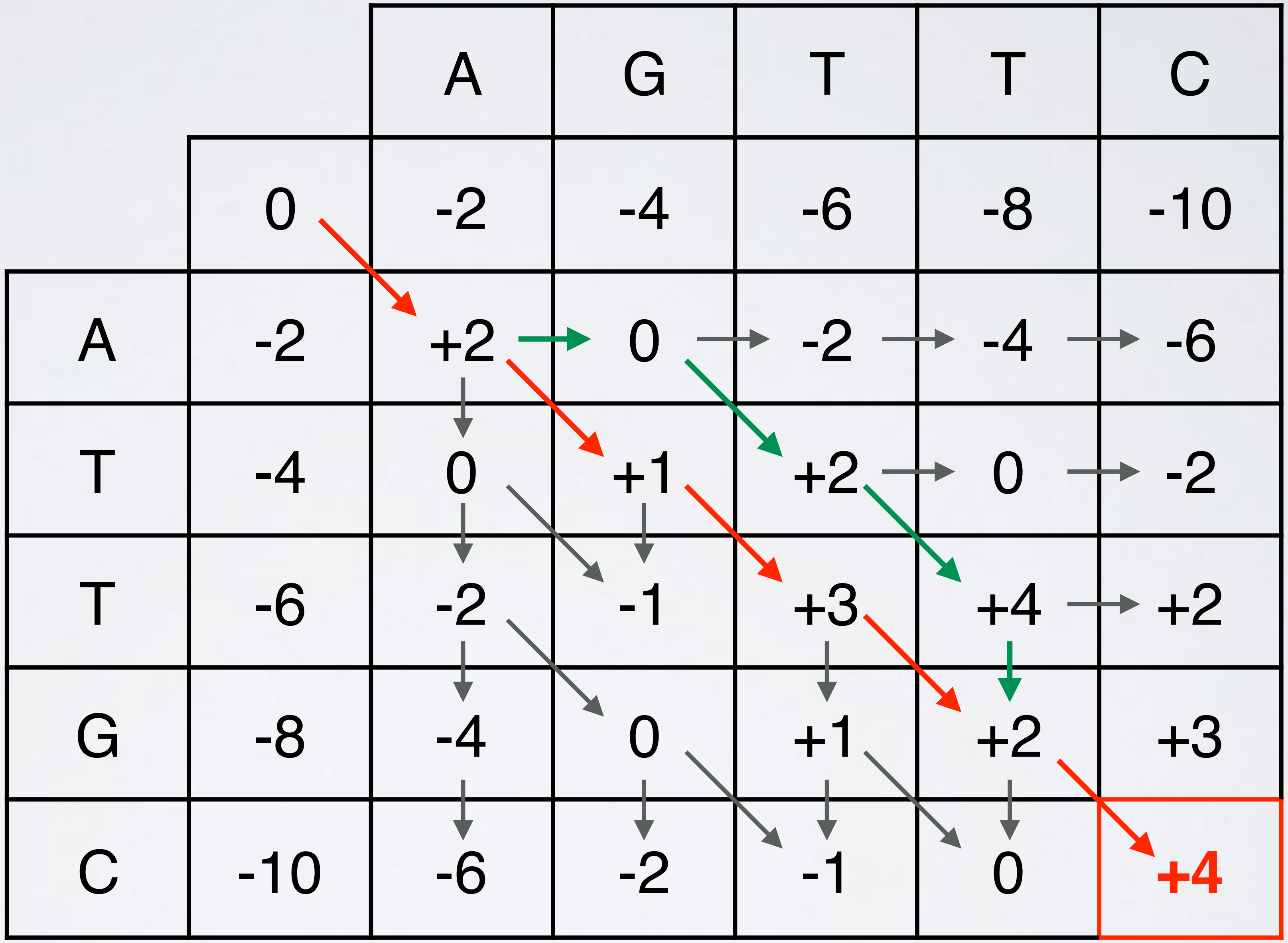
- *“More time to finish the assignment”*
- *“The [NCBI] sites were so slow”*
- *“More time with HMMER would be helpful”*
- *“Very nice lab”*

Q18: NW DYNAMIC PROGRAMMING

Match: +2
Mismatch: -1
Gap: -2

```
A T T G C
I   I   I
A G T T C

A - T T G C
I   I I   I
A G T T - C
```



		A	G	T	T	C
	0	-2	-4	-6	-8	-10
A	-2	+2	0	-2	-4	-6
T	-4	0	+1	+2	0	-2
T	-6	-2	-1	+3	+4	+2
G	-8	-4	0	+1	+2	+3
C	-10	-6	-2	-1	0	**+4**

THIS WEEK'S HOMEWORK

☑ Check out the "**Background Reading**" material online:

- Achievements & Challenges in Structural Bioinformatics
- Protein Structure Prediction
- Biomolecular Simulation
- Computational Drug Discovery

☑ Complete the **lecture 1.3 homework questions**:
http://tinyurl.com/bioinf525-quiz3

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

… A hybrid of biology and computer science

“Bioinformatics is the application of <u>computers</u> to the collection, archiving, organization, and analysis of <u>biological data</u>.”

Bioinformatics is computer aided biology!

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

Bioinformatics is computer aided biology!

Goal: Data to Knowledge

So what is **structural bioinformatics**?

So what is **structural bioinformatics**?

… computer aided structural biology!

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

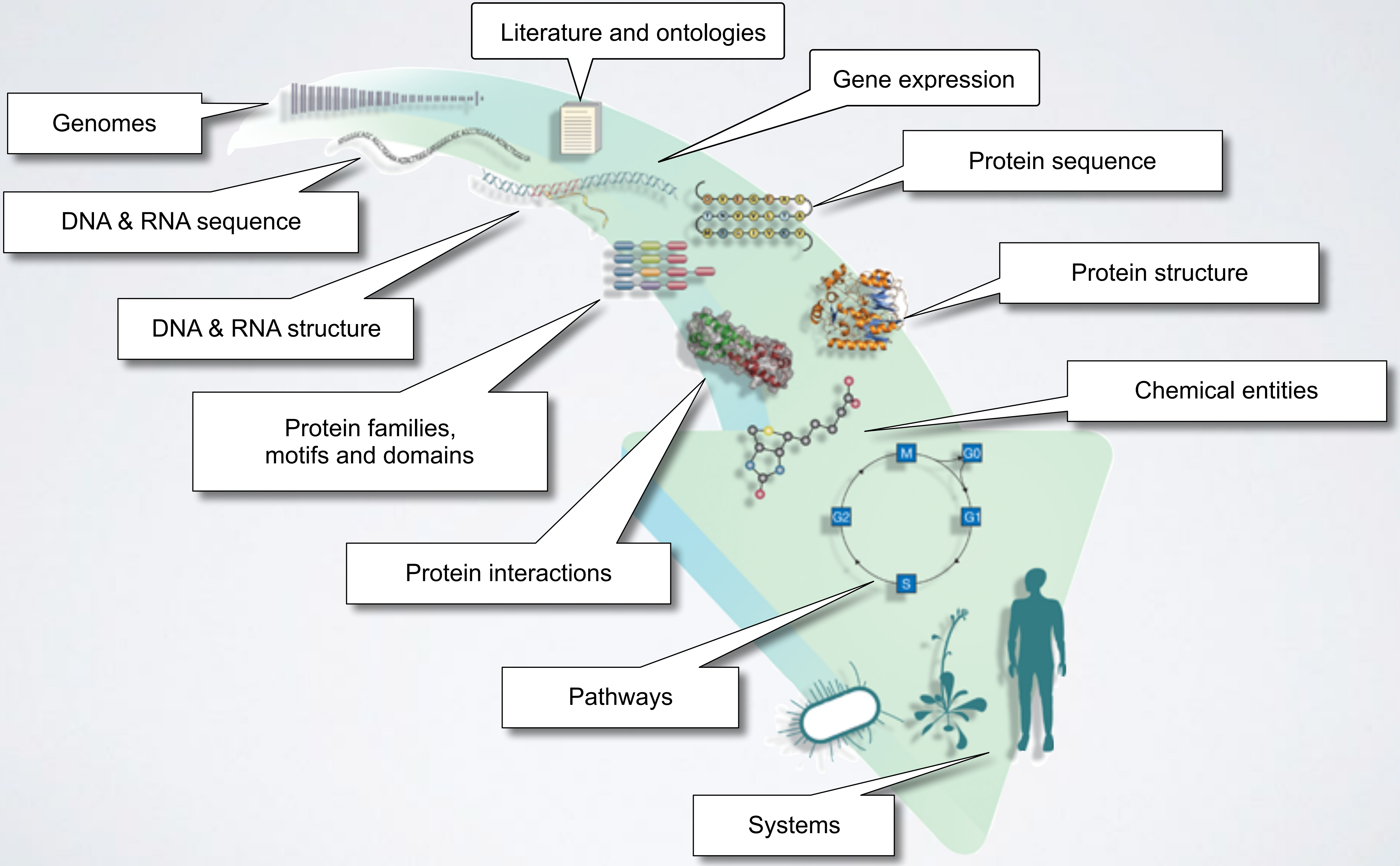
Why should we care?

Why should we care?

Because biomolecules are “nature’s robots”

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

BIOINFORMATICS DATA



Literature and ontologies

Genomes

Gene expression

Protein sequence

DNA & RNA sequence

Protein structure

DNA & RNA structure

Chemical entities

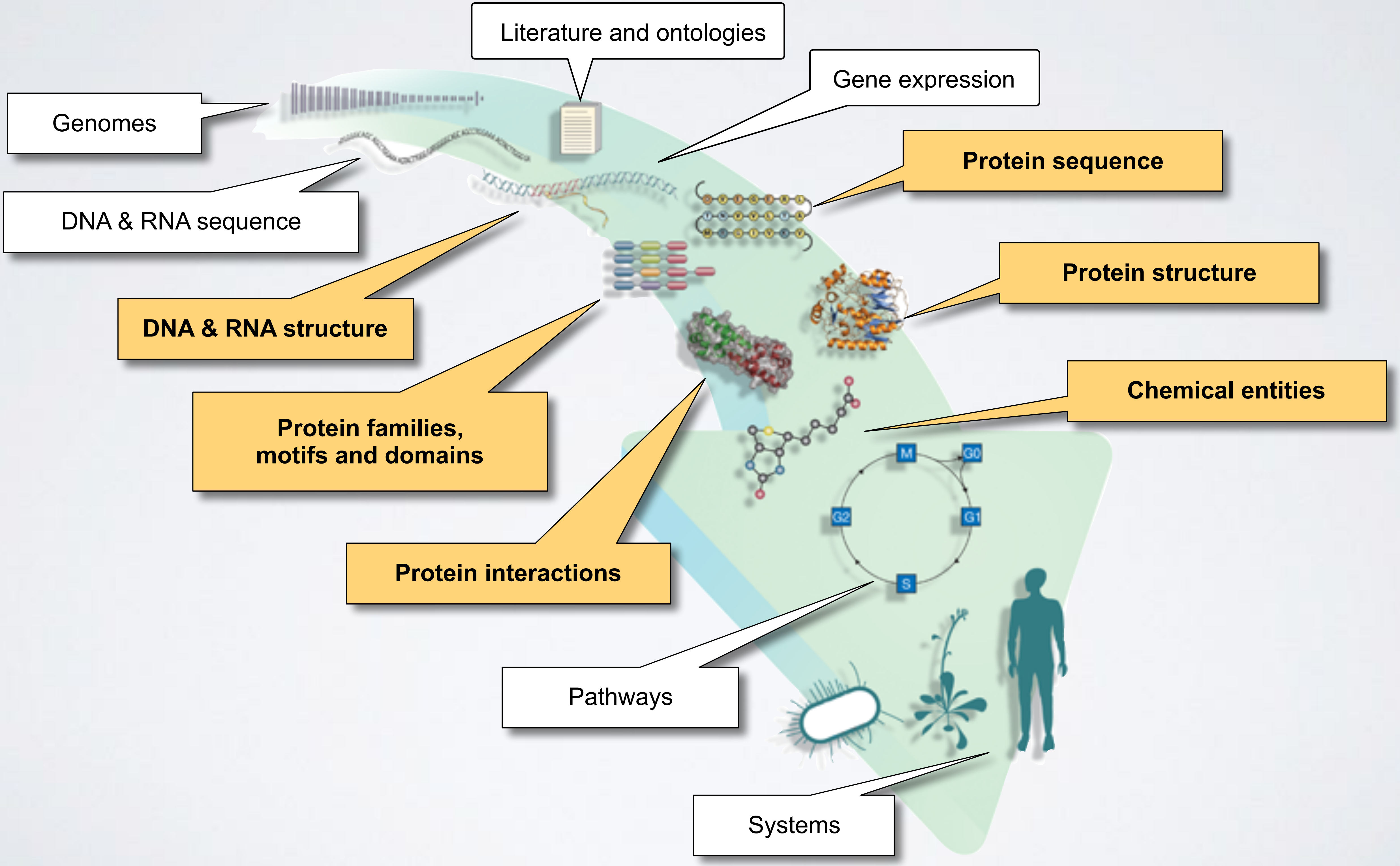
Protein families, motifs and domains

Protein interactions

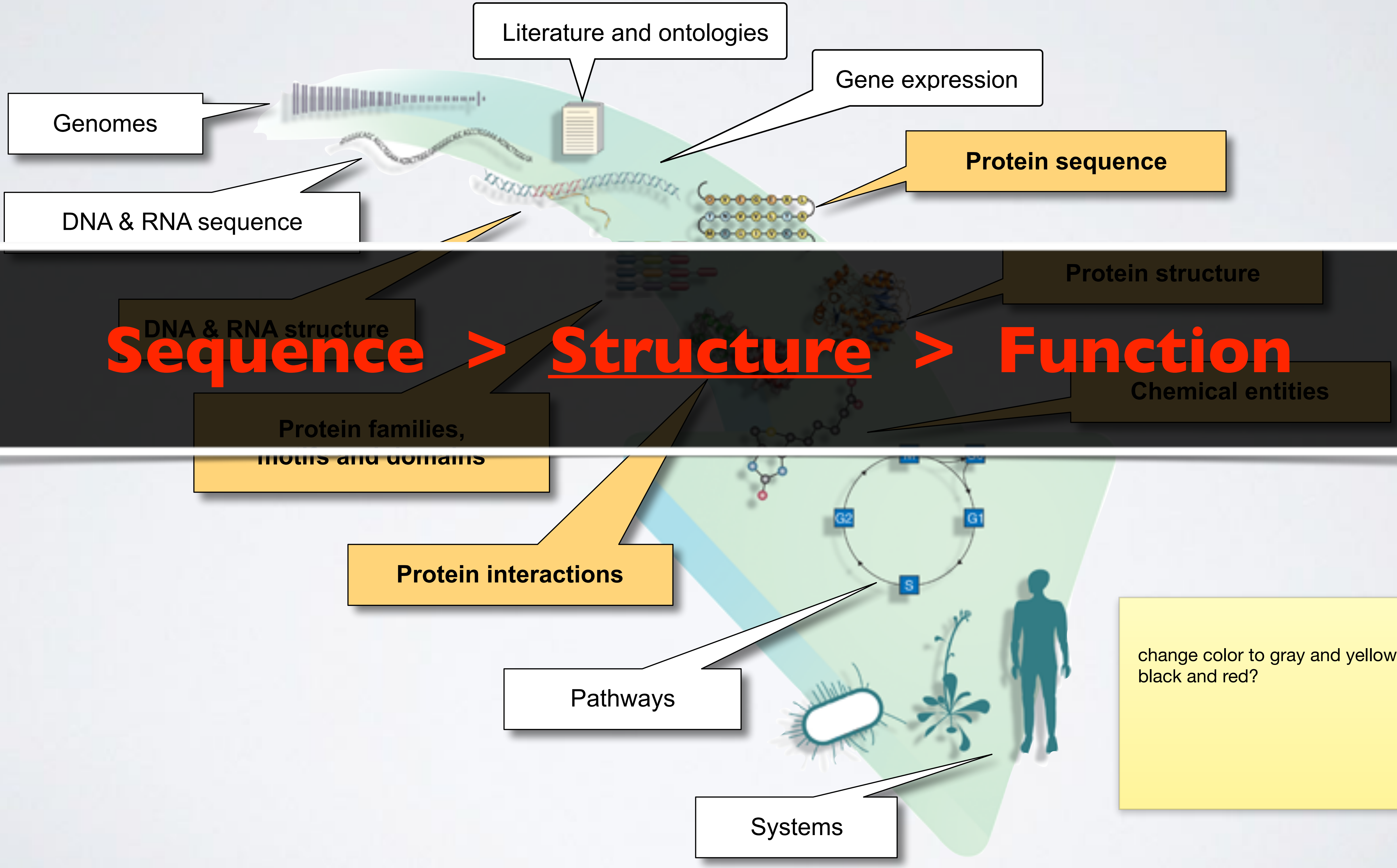
Pathways

Systems

STRUCTURAL DATA IS CENTRAL



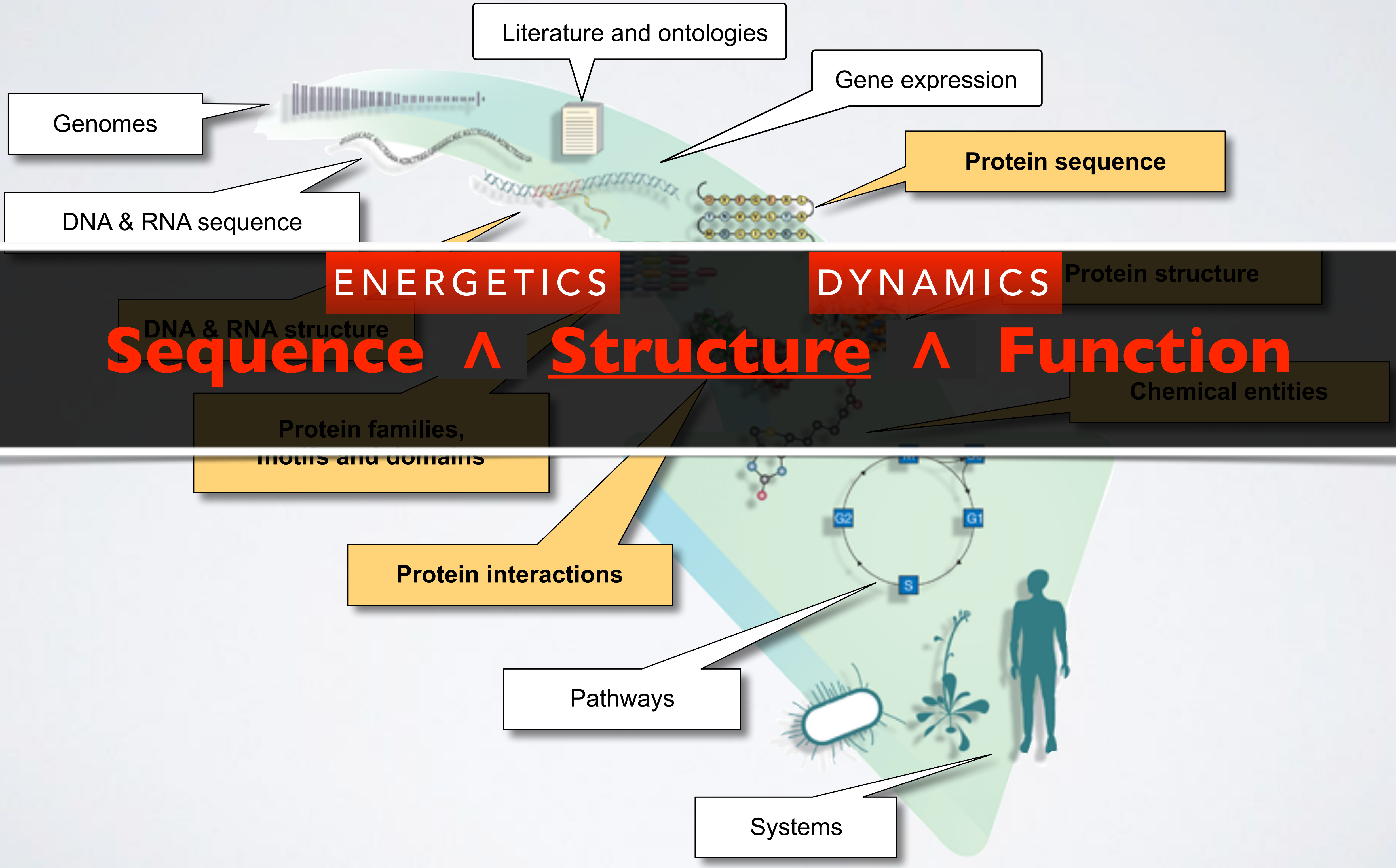
STRUCTURAL DATA IS CENTRAL



Sequence > Structure > Function

change color to gray and yellow from black and red?

STRUCTURAL DATA IS CENTRAL



Literature and ontologies

Genomes

Gene expression

Protein sequence

DNA & RNA sequence

Protein structure

DNA & RNA structure

Chemical entities

Protein families, motifs and domains

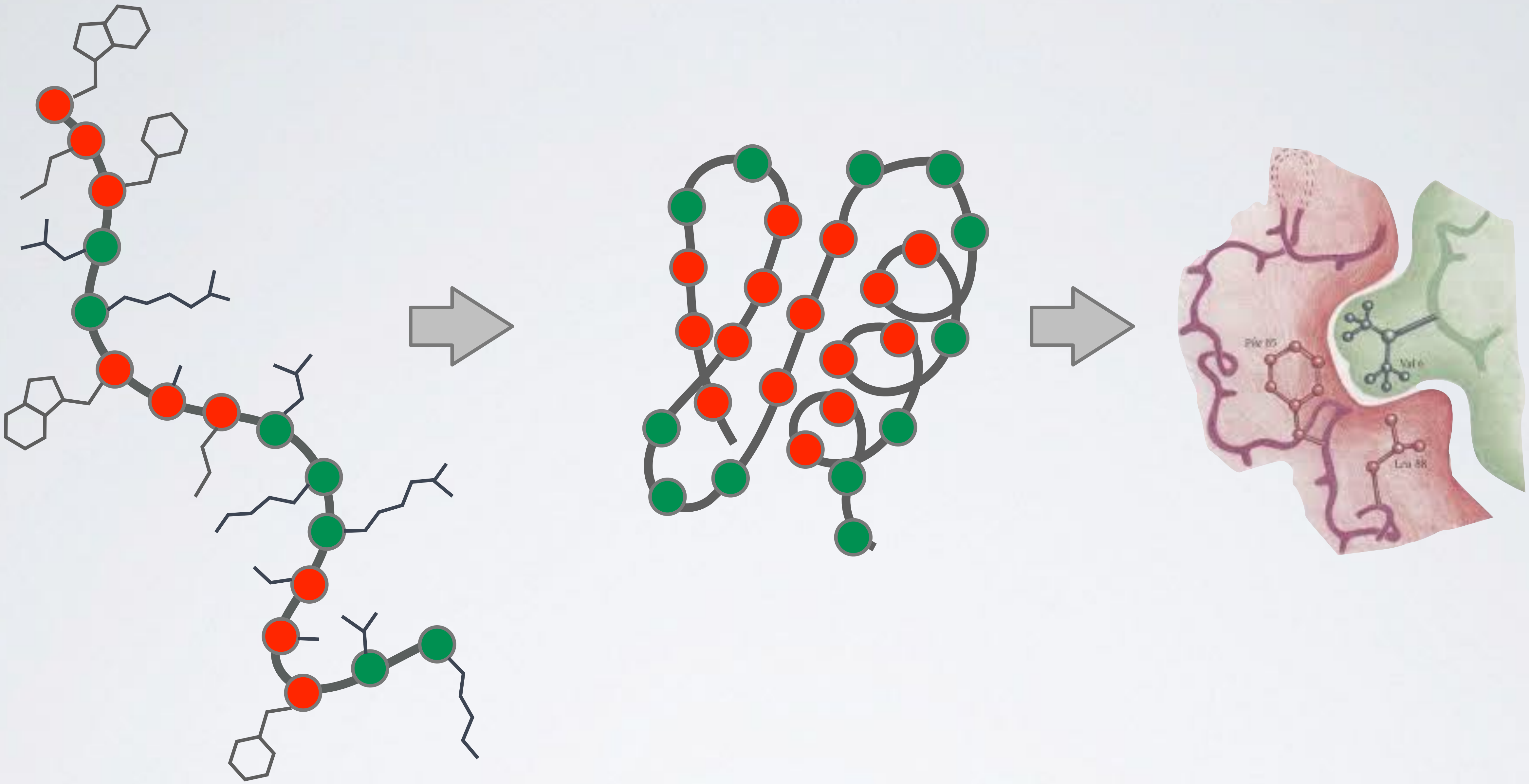
Protein interactions

Pathways

Systems

ENERGETICS DYNAMICS

Sequence ∧ Structure ∧ Function



Sequence

- Unfolded chain of amino acid chain
- Highly mobile
- Inactive

Structure

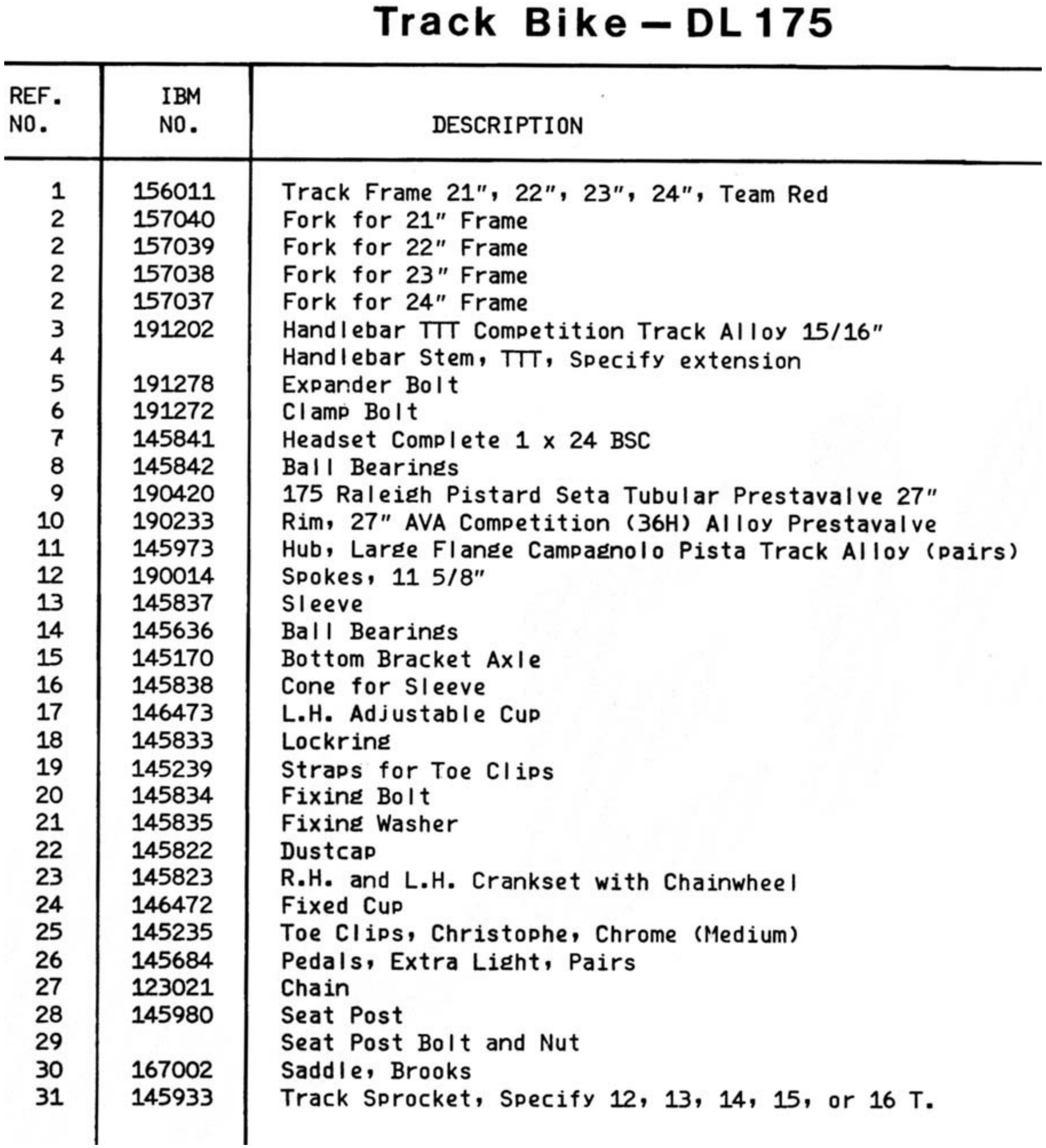
- Ordered in a precise 3D arrangment
- Stable but dynamic

Function

- Active in specific "conformations"
- Specific associations & precise reactions

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

Genomics is a great start ….

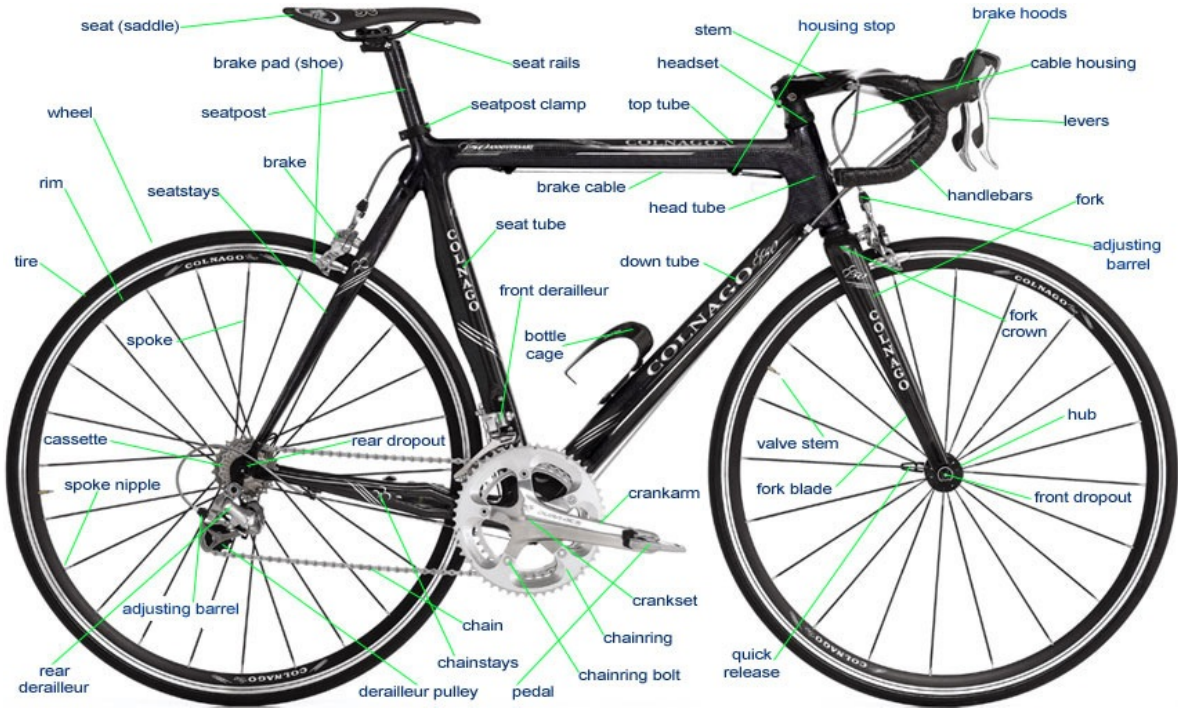


Track Bike — DL 175

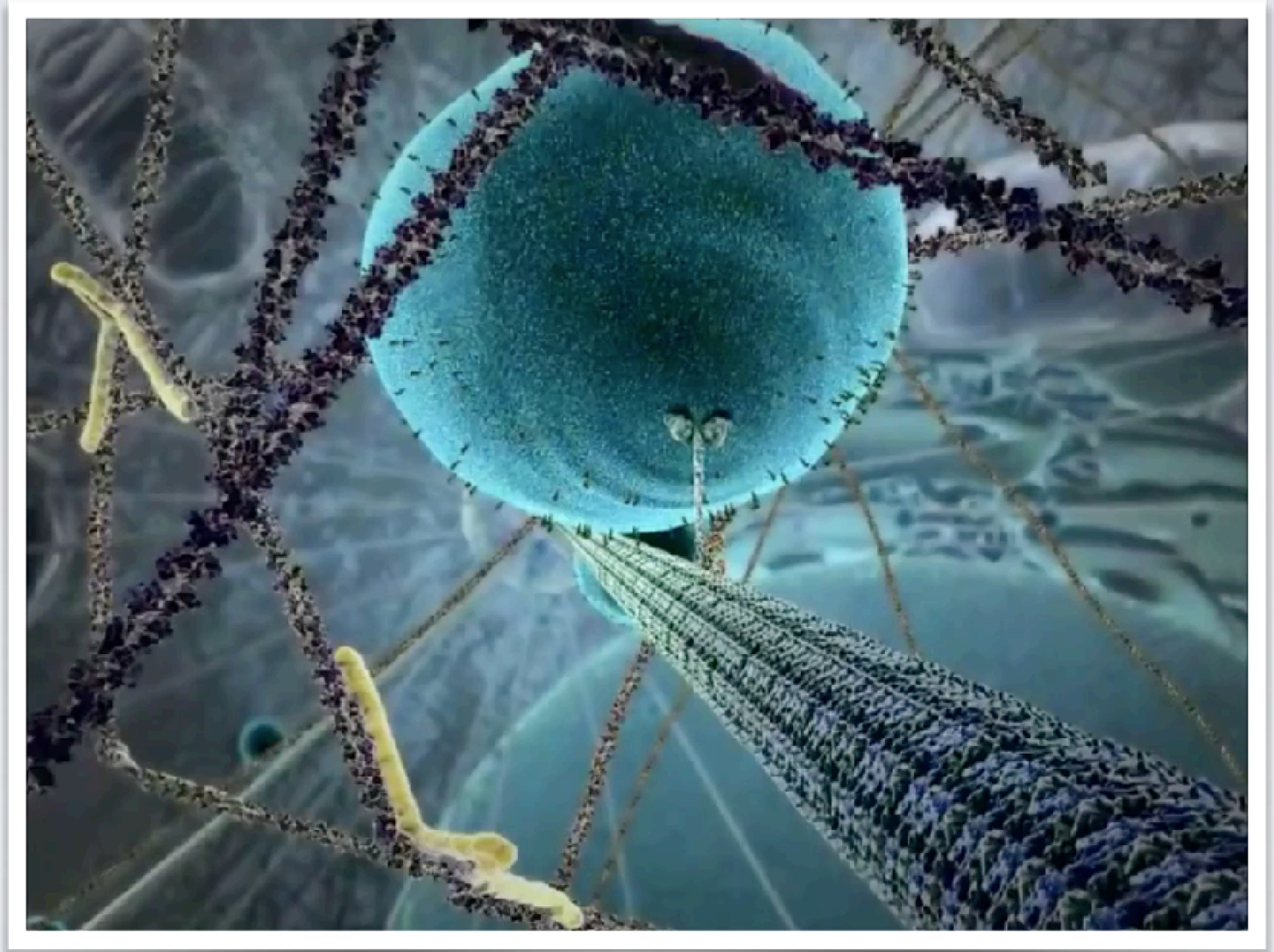
REF. NO.	IBM NO.	DESCRIPTION
1	156011	Track Frame 21″, 22″, 23″, 24″, Team Red
2	157040	Fork for 21″ Frame
2	157039	Fork for 22″ Frame
2	157038	Fork for 23″ Frame
2	157037	Fork for 24″ Frame
3	191202	Handlebar TTT Competition Track Alloy 15/16″
4		Handlebar Stem, TTT, Specify extension
5	191278	Expander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

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

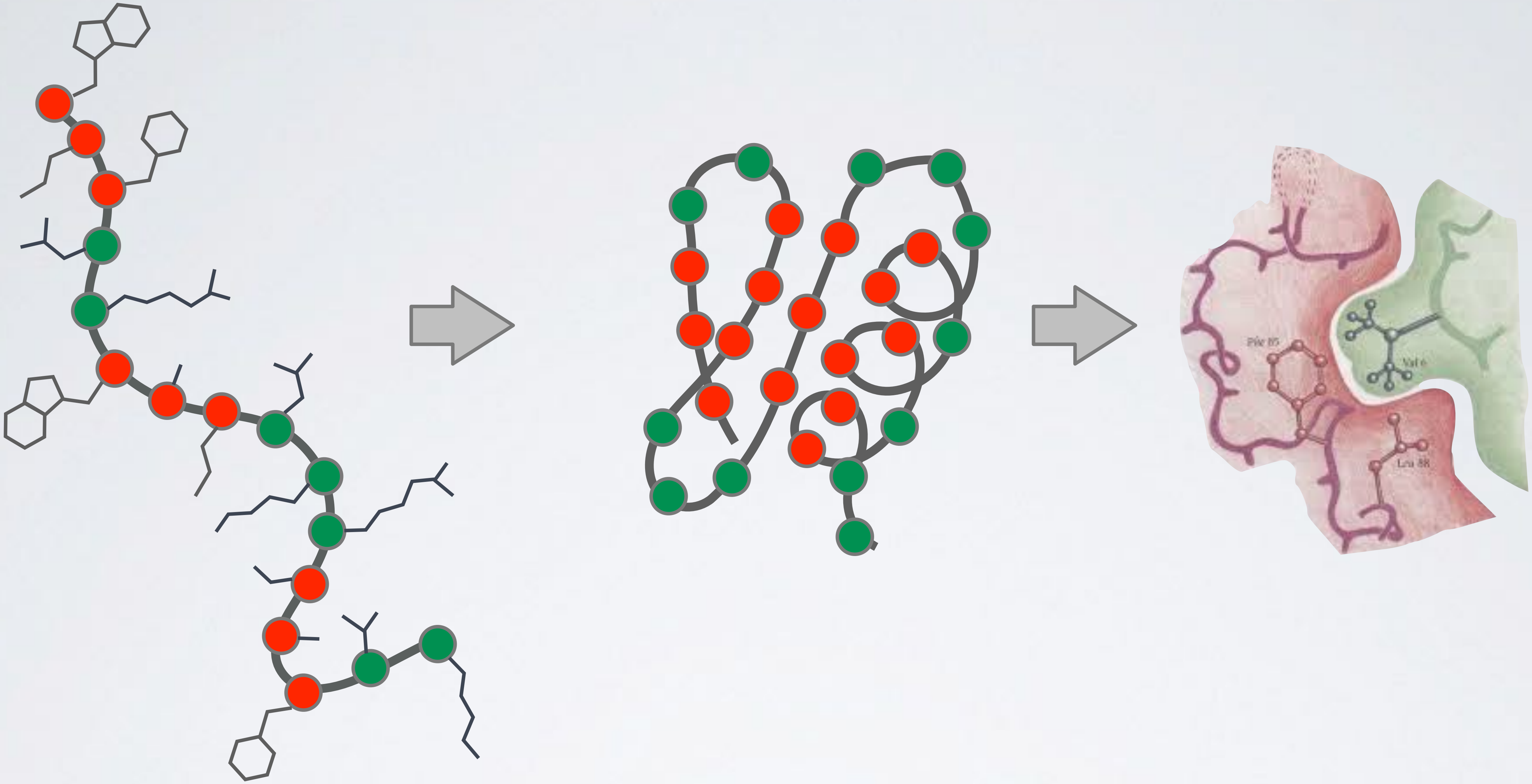
… but not the end



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



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



Sequence

- Unfolded chain of amino acid chain
- Highly mobile
- Inactive

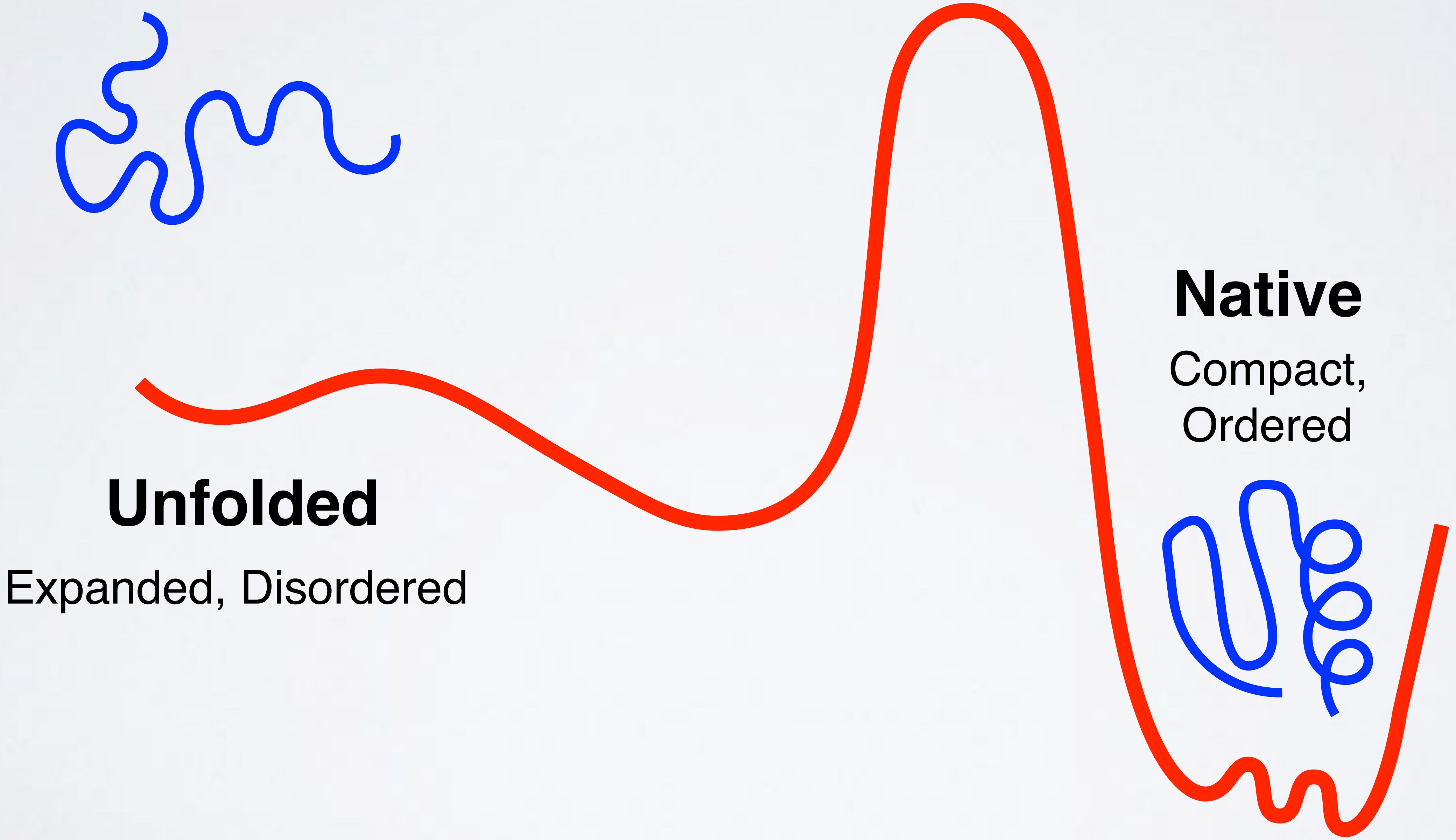
Structure

- Ordered in a precise 3D arrangment
- Stable but dynamic

Function

- Active in specific "conformations"
- Specific associations & precise reactions

KEY CONCEPT: ENERGY LANDSCAPE



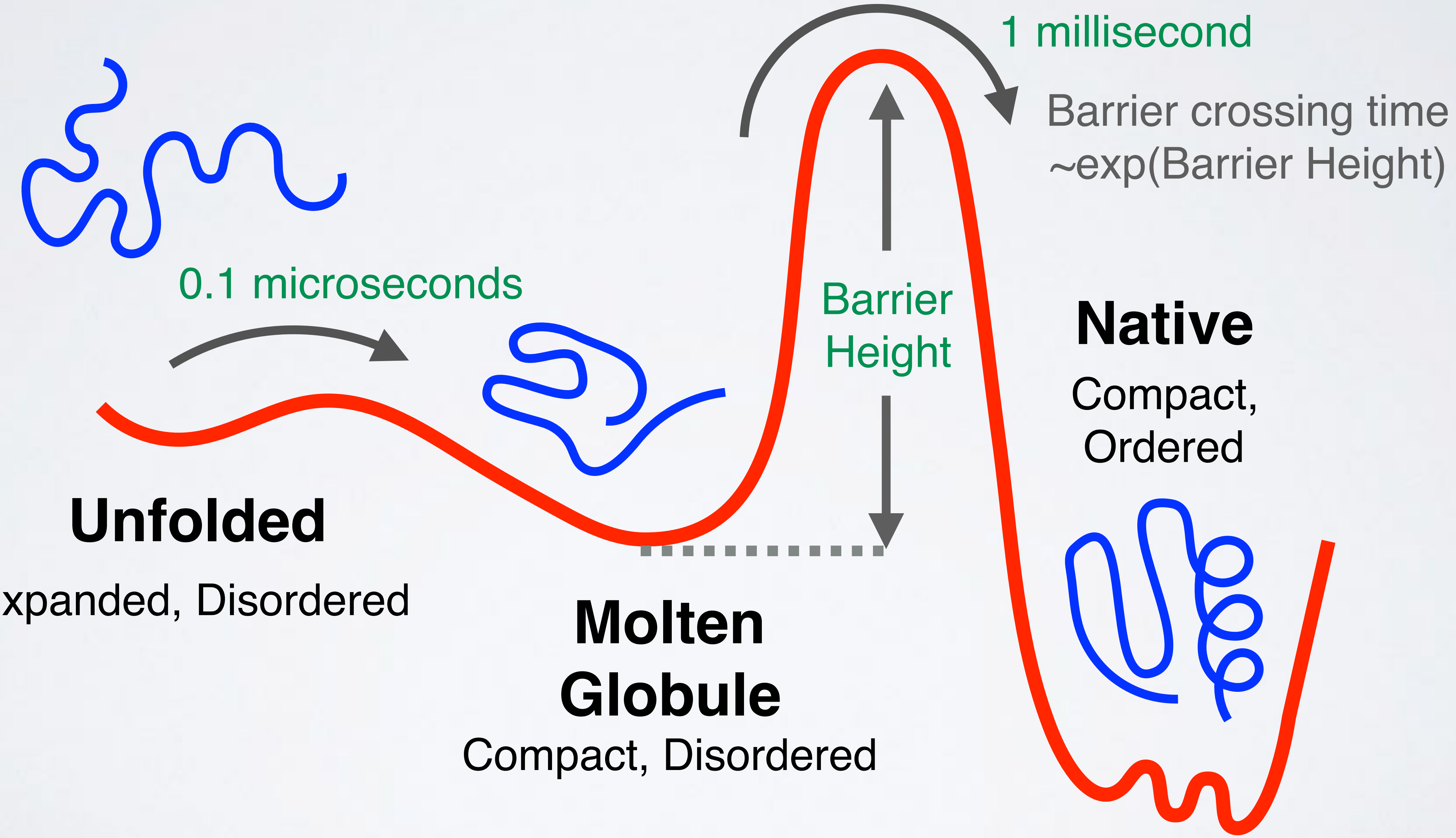
Unfolded

Expanded, Disordered

Native

Compact,
Ordered

KEY CONCEPT: ENERGY LANDSCAPE



0.1 microseconds

1 millisecond

Barrier crossing time ~exp(Barrier Height)

Barrier Height

Unfolded

Expanded, Disordered

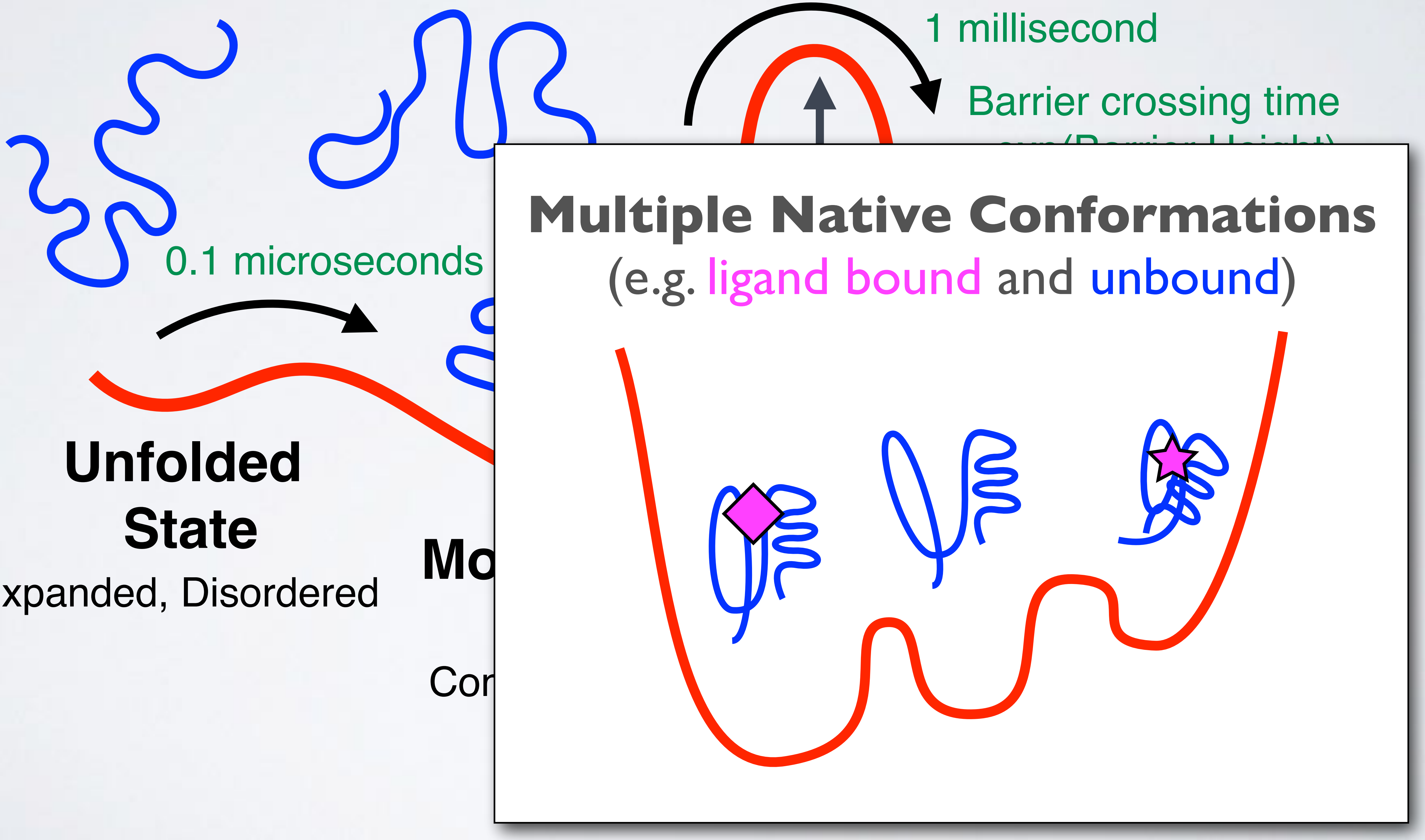
Molten Globule

Compact, Disordered

Native

Compact, Ordered

KEY CONCEPT: **ENERGY LANDSCAPE**



1 millisecond

Barrier crossing time

exp(Barrier Height)

0.1 microseconds

Unfolded State

Expanded, Disordered

Mo

Cor

Multiple Native Conformations

(e.g. ligand bound and unbound)

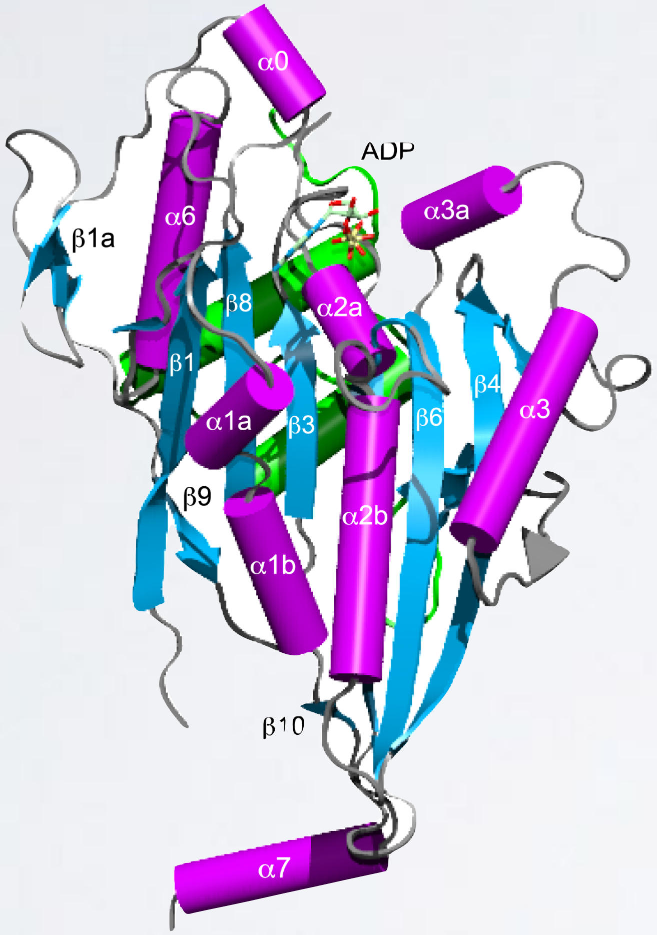
OUTLINE:

- **Overview of structural bioinformatics**
  - Major motivations, goals and challenges
- **Fundamentals of protein structure**
  - Composition, form, forces and dynamics
- **Representing and interpreting protein structure**
  - Modeling energy as a function of structure
- **Example application areas**
  - Predicting functional dynamics & drug discovery

OUTLINE:

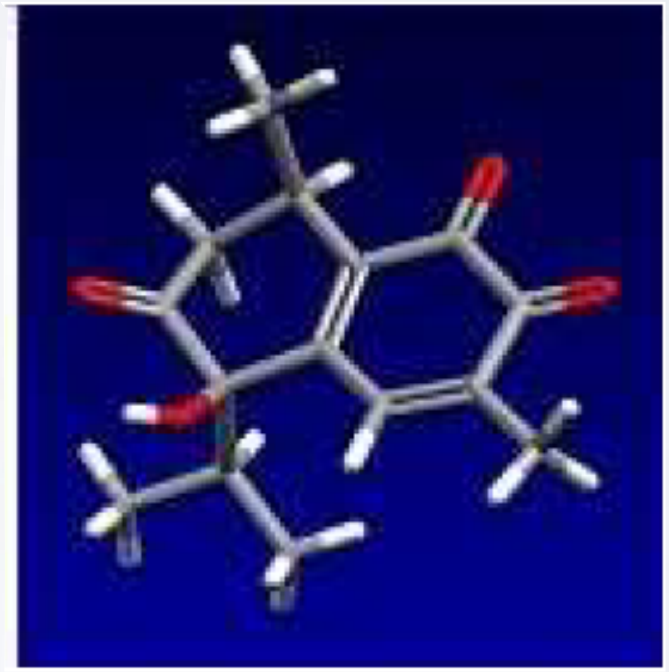
- **Overview of structural bioinformatics**
  - Major motivations, goals and challenges
- **Fundamentals of protein structure**
  - Composition, form, forces and dynamics
- **Representing and interpreting protein structure**
  - Modeling energy as a function of structure
- **Example application areas**
  - Predicting functional dynamics & drug discovery

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



Protein
(PDB)

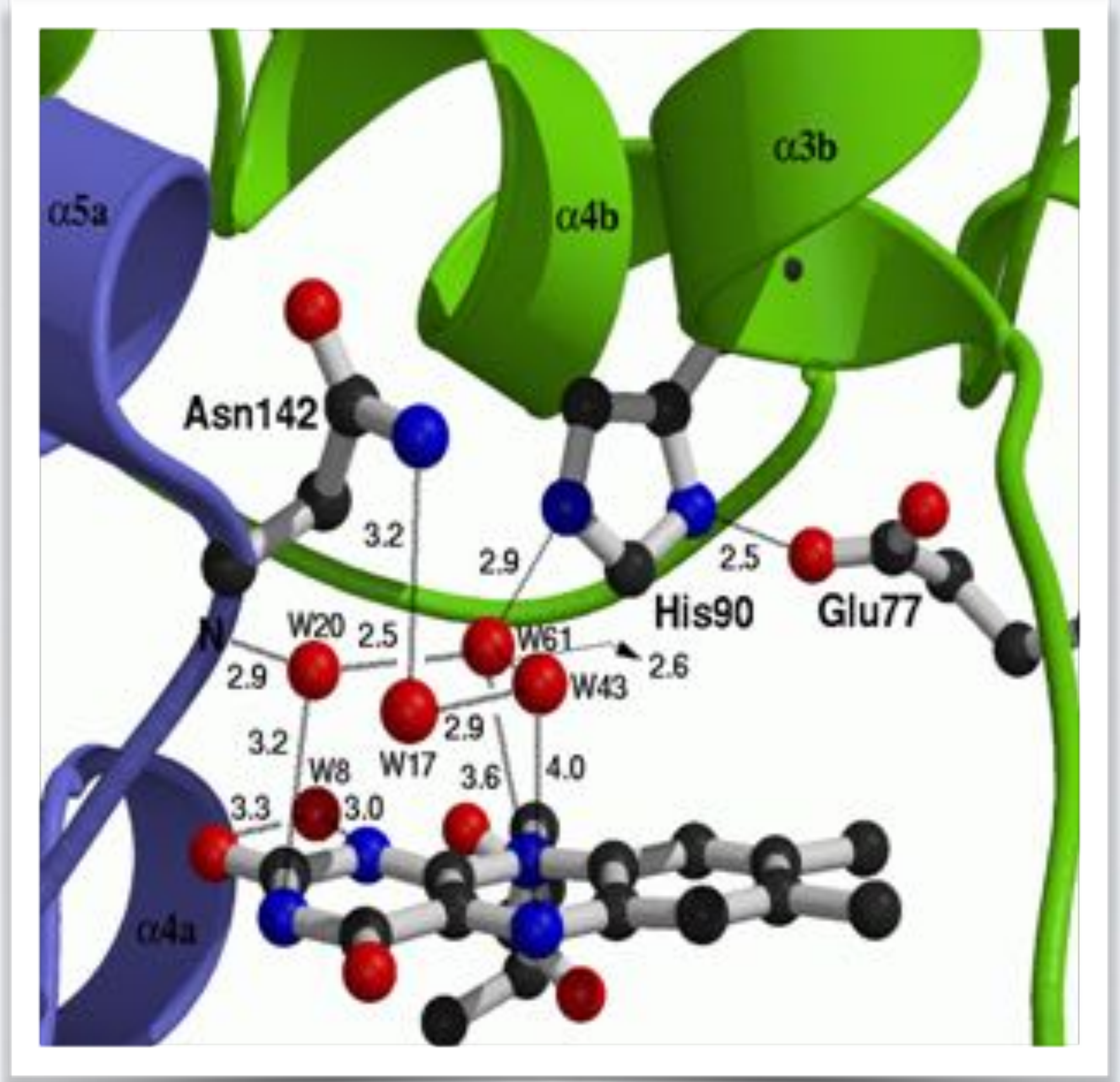
DNA
(NDB)



Small Molecules
(CCDB)

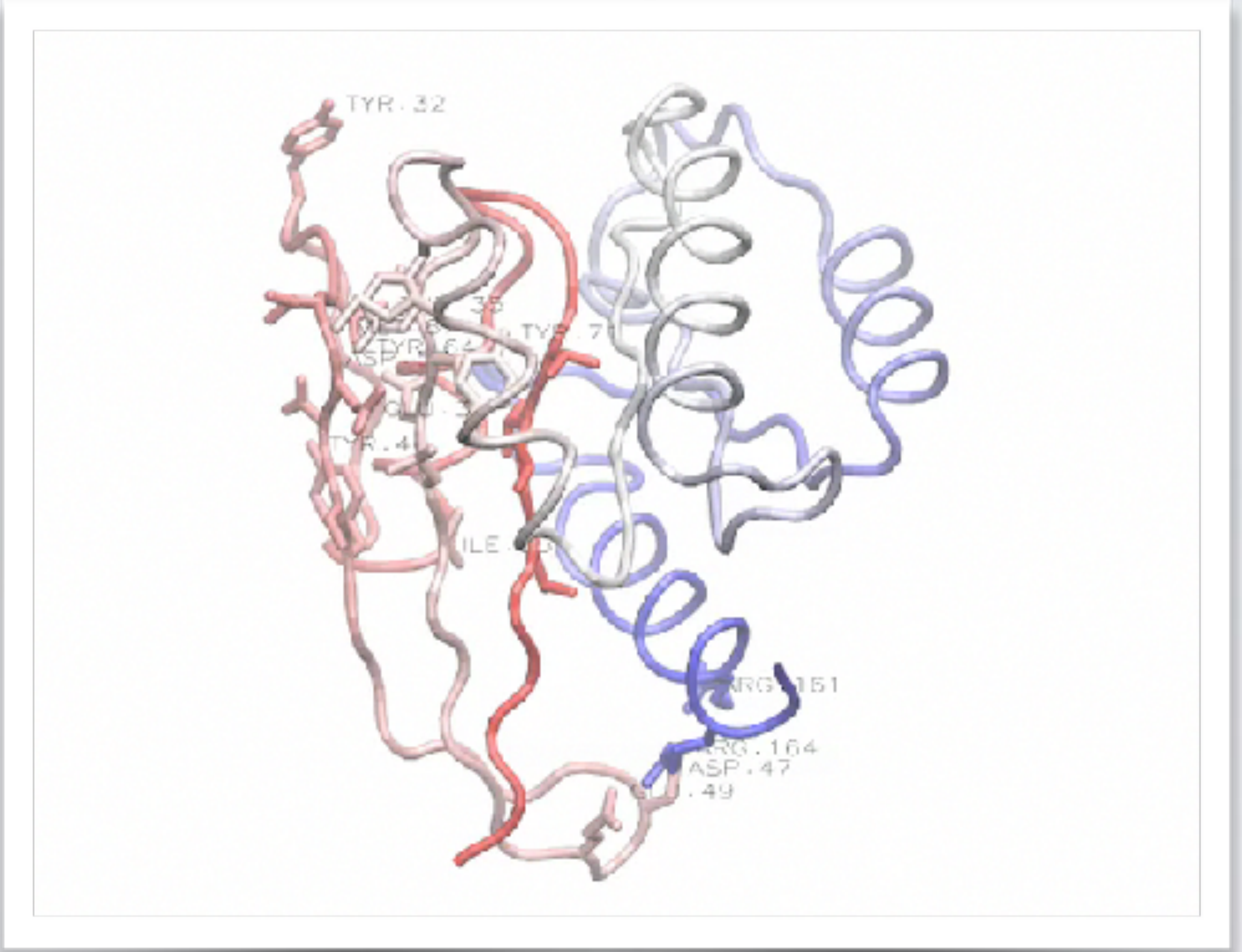
Motivation 1:
Detailed understanding of molecular interactions

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



Motivation 1:
Detailed understanding of molecular interactions

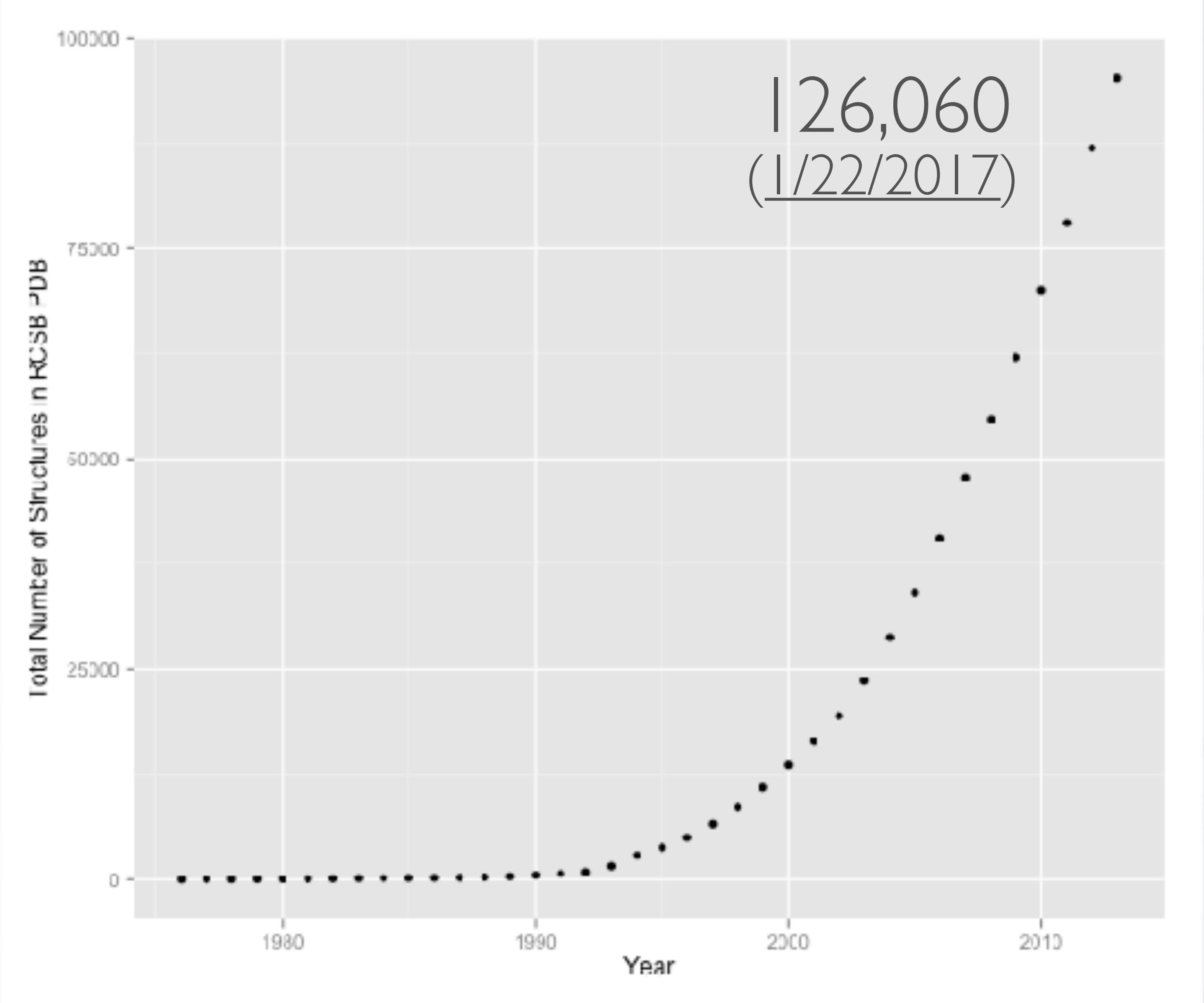
Computational modeling can provide detailed insight into functional interactions, their regulation and potential consequences of perturbation.



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

Motivation 2:
Lots of structural data is becoming available

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



126,060
(1/22/2017)

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

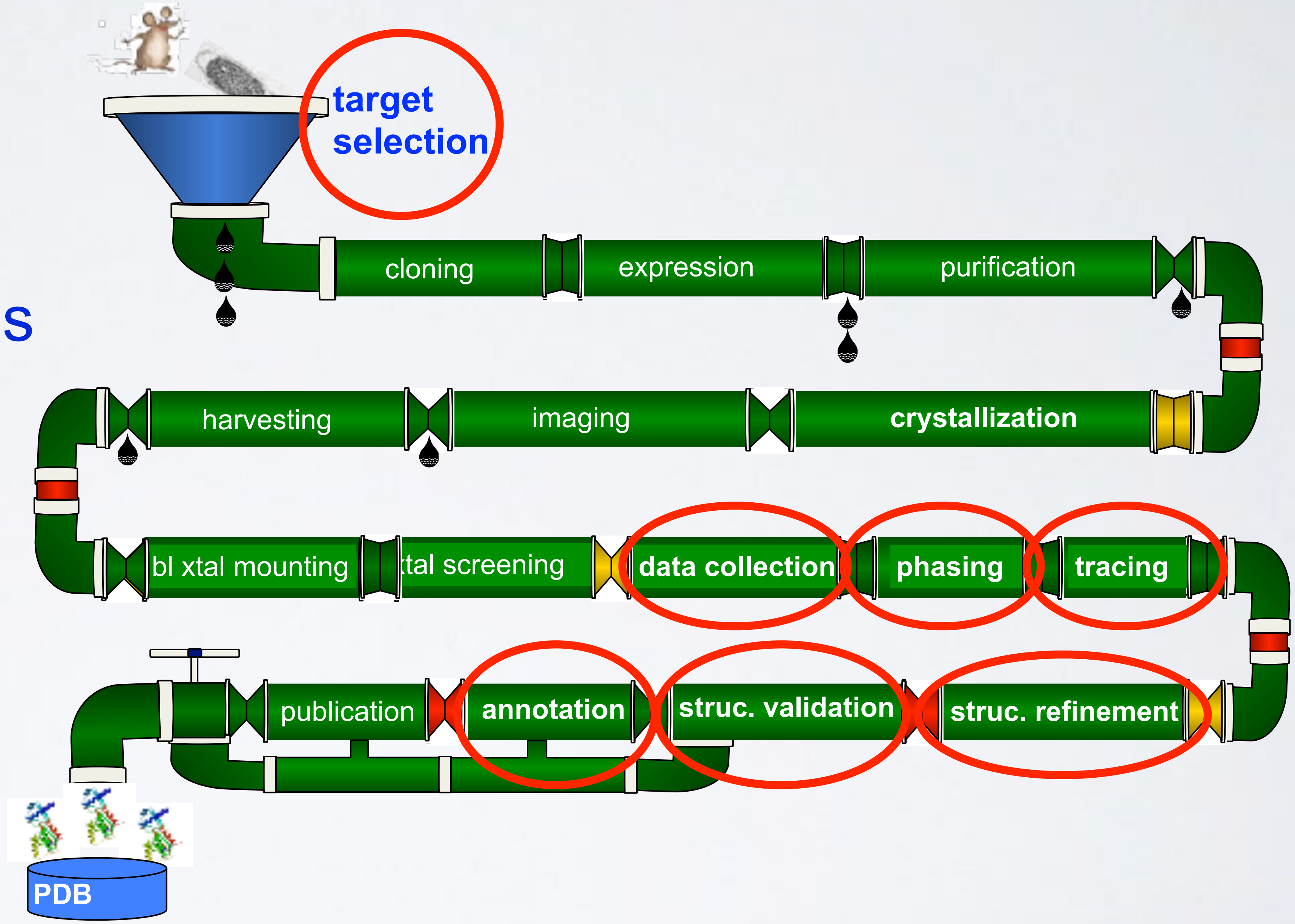
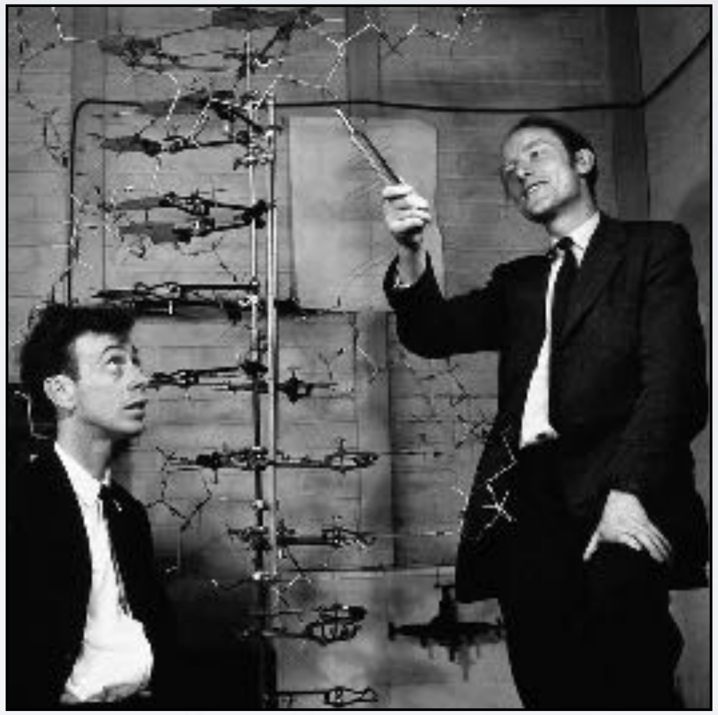
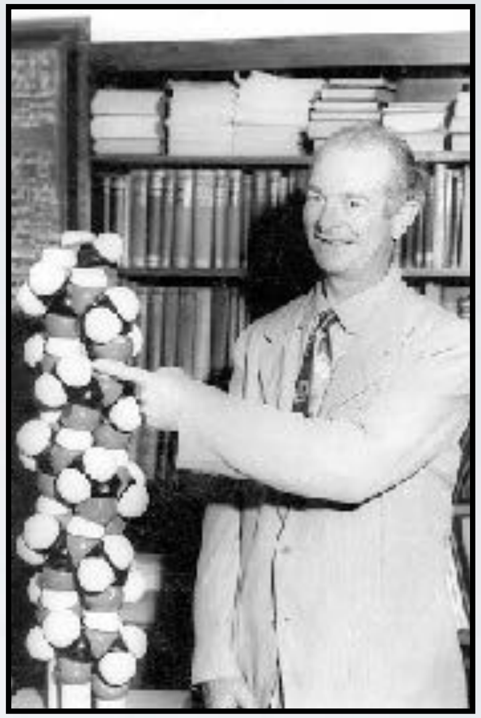


Image Credit: "Structure determination assembly line" Adam Godzik





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

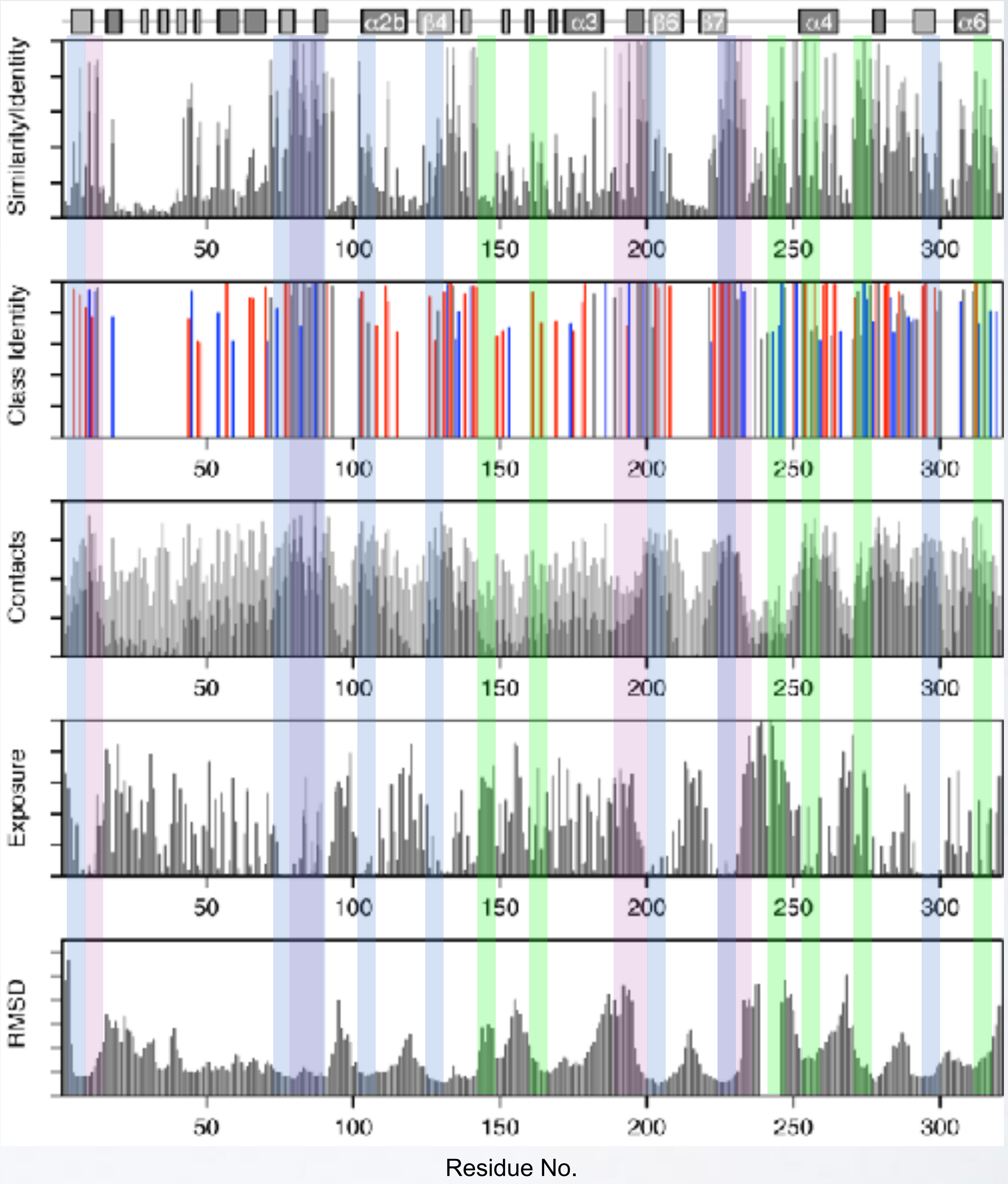
- Structure allows identification of more distant evolutionary relationships

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)

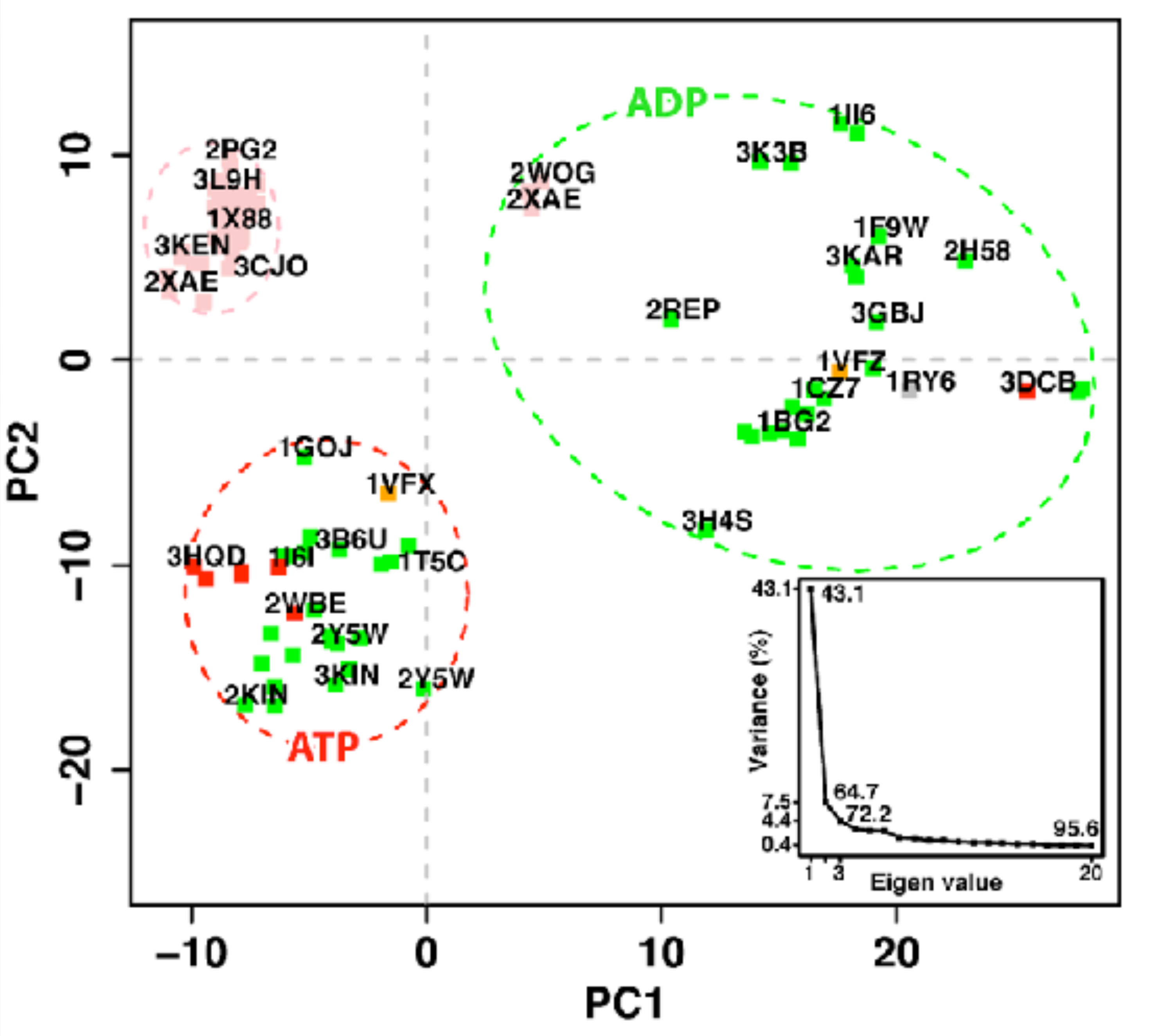
Goals:

- Analysis
- Visualization
- Comparison
- Prediction
- Design

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

Goals:

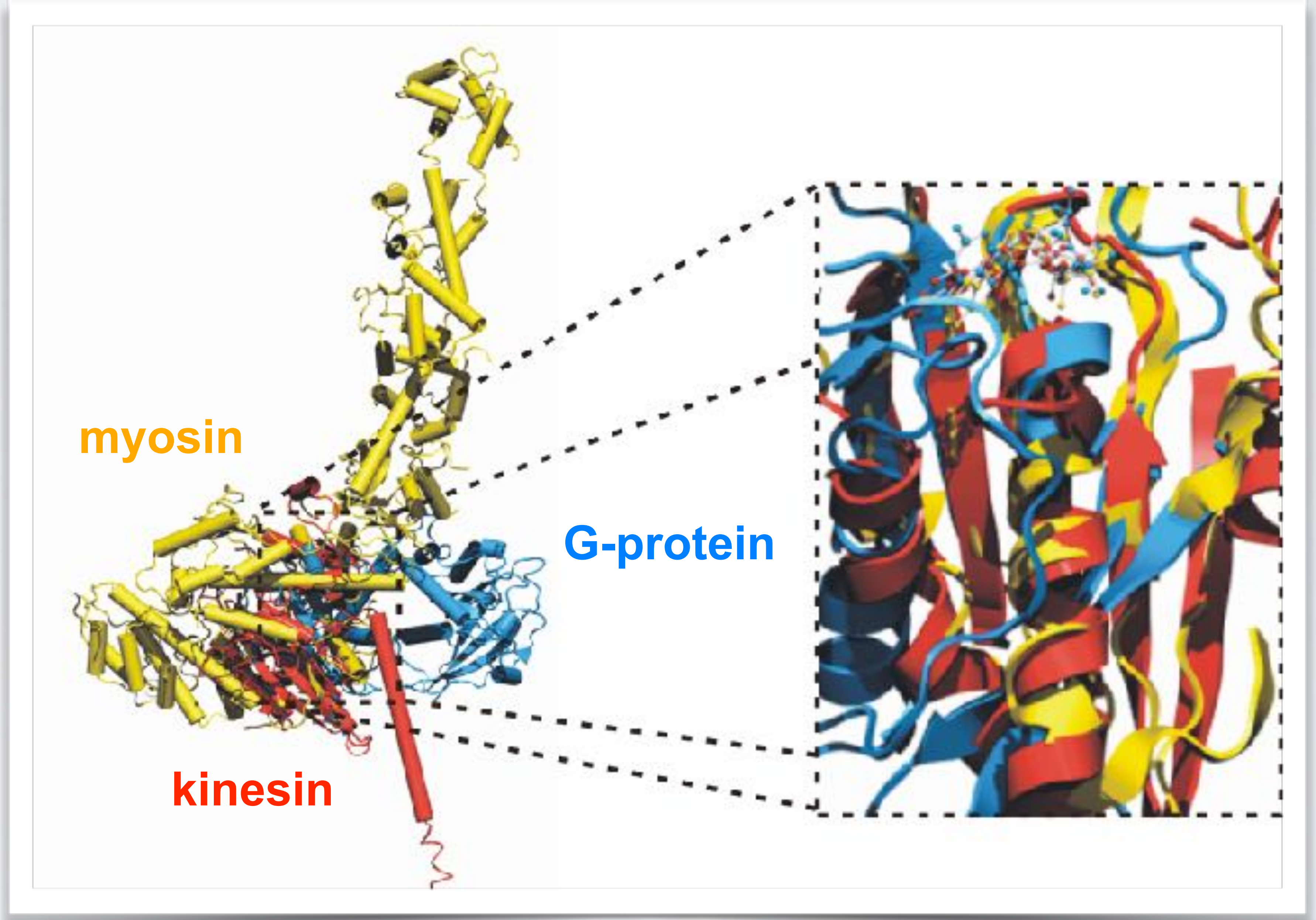
- Analysis
- Visualization
- Comparison
- Prediction
- Design



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

Goals:

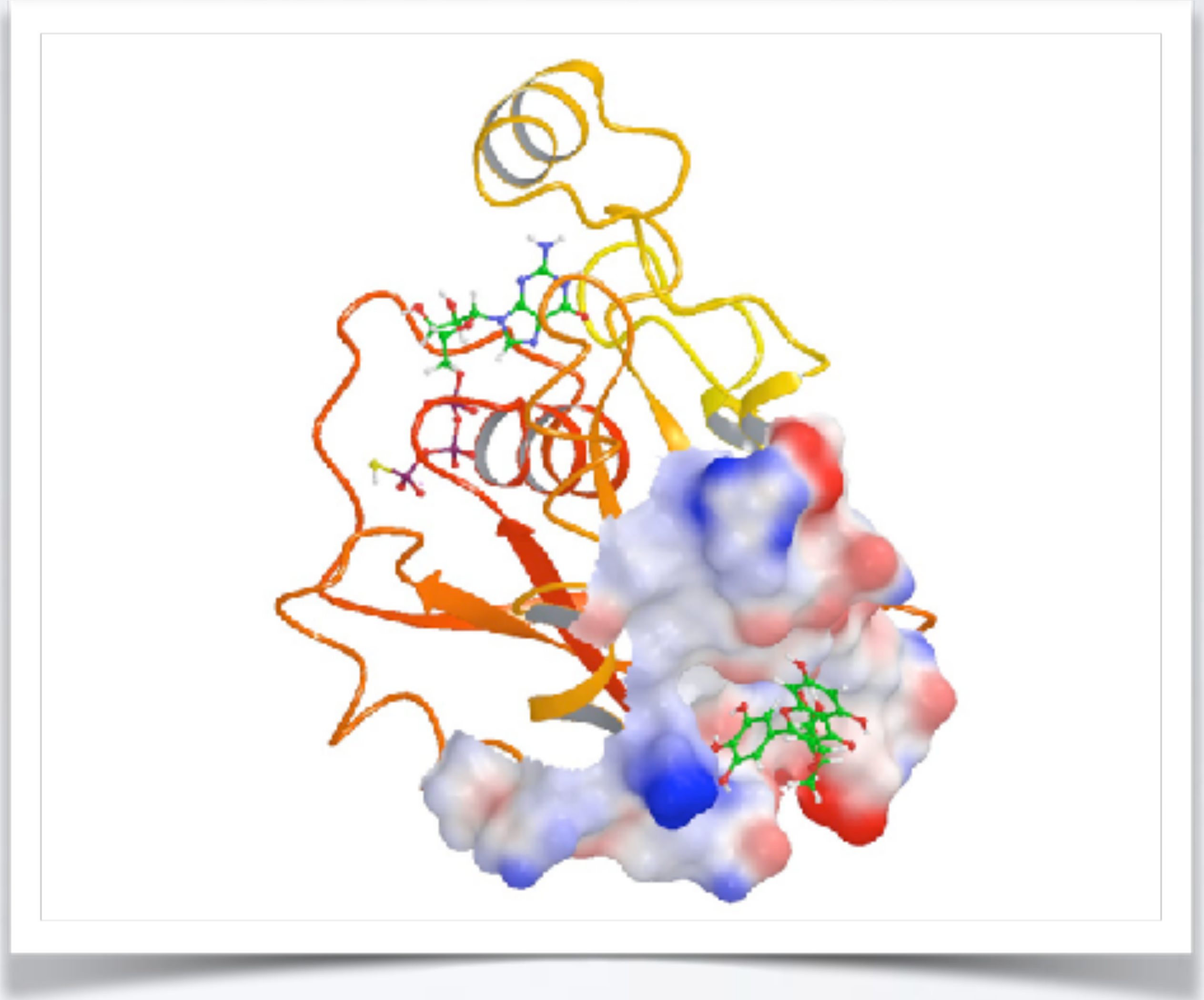
- Analysis
- Visualization
- Comparison
- Prediction
- Design



Grant *et al.* unpublished

Goals:

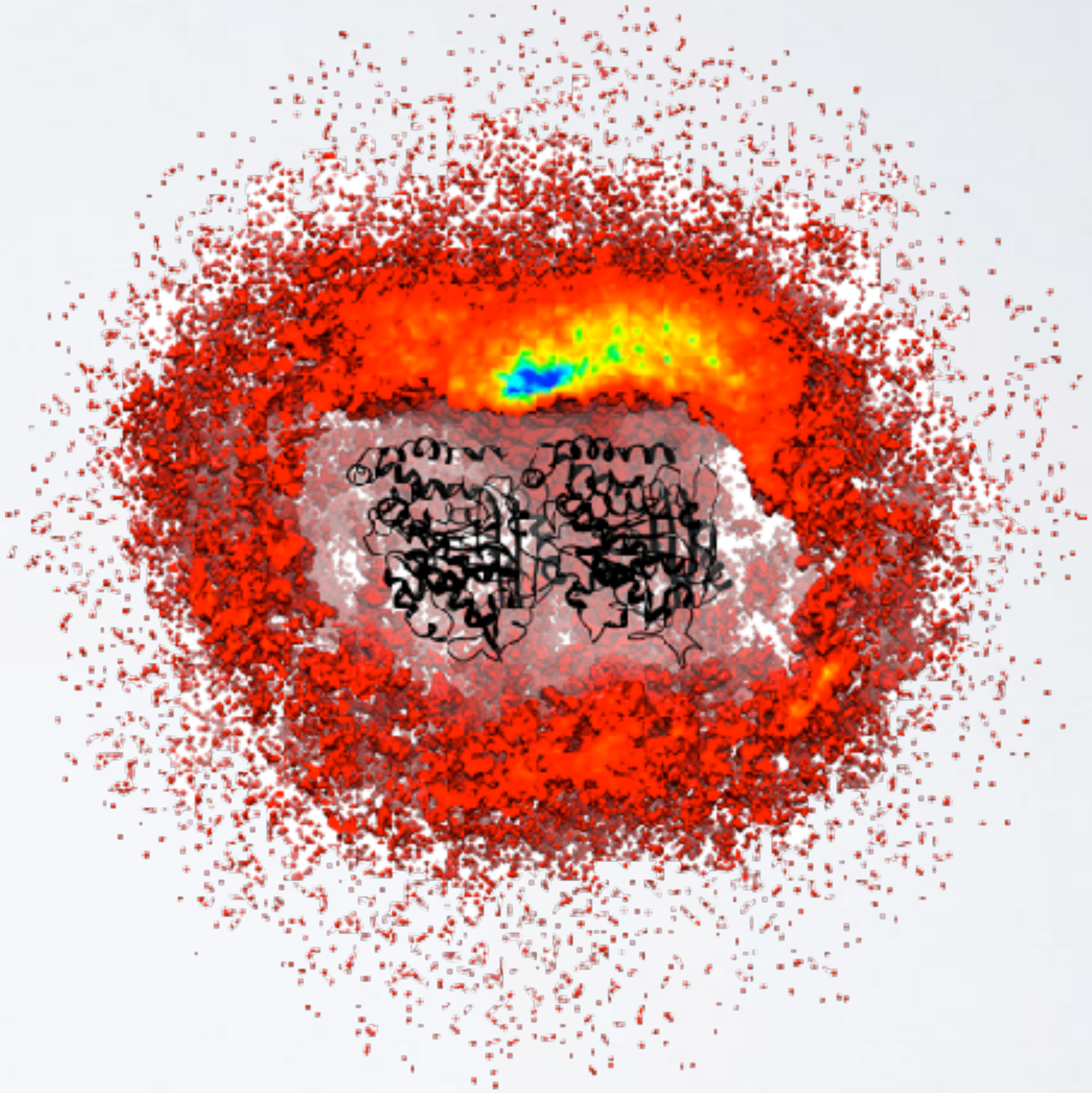
- Analysis
- Visualization
- Comparison
- Prediction
- Design



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

NEXT UP:

- **Overview of structural bioinformatics**
  - Major motivations, goals and challenges

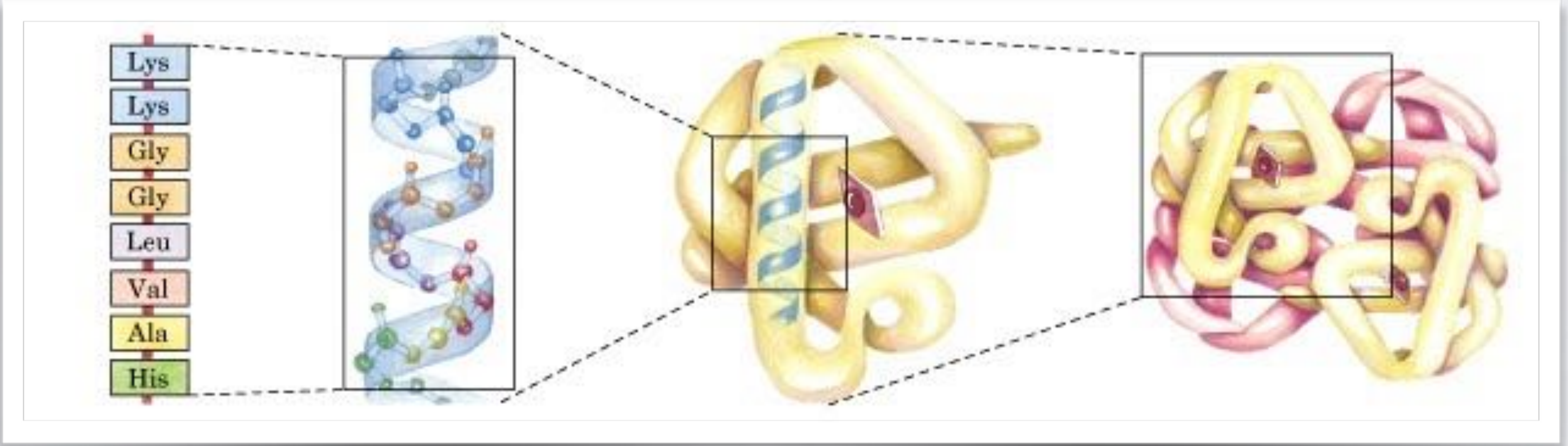
- **Fundamentals of protein structure**
  - Composition, form, forces and dynamics

- **Representing and interpreting protein structure**
  - Modeling energy as a function of structure

- **Example application areas**
  - Predicting functional dynamics & drug discovery

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

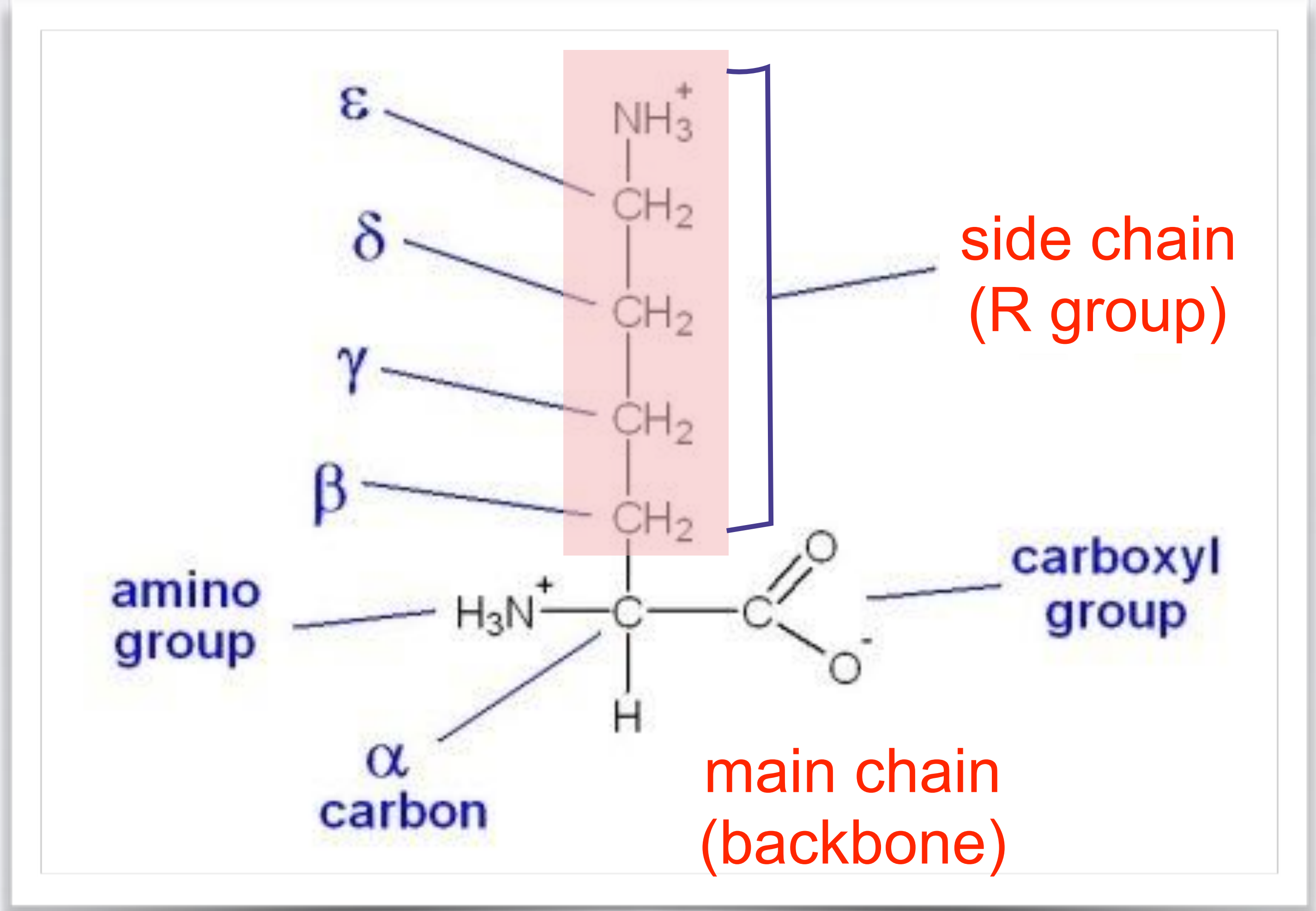
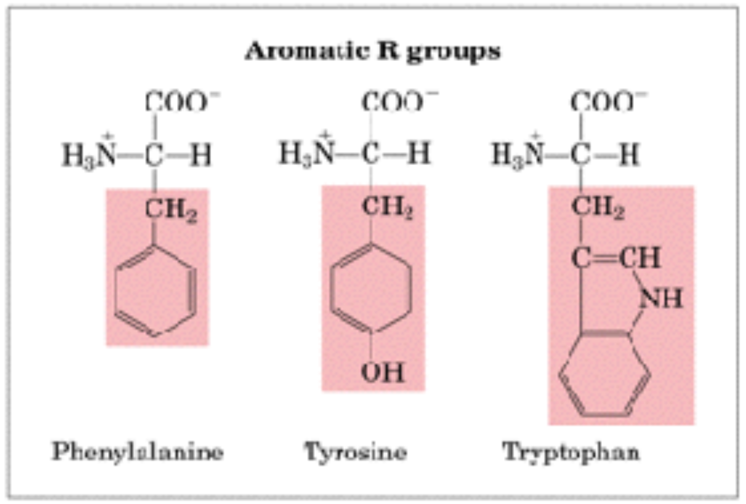
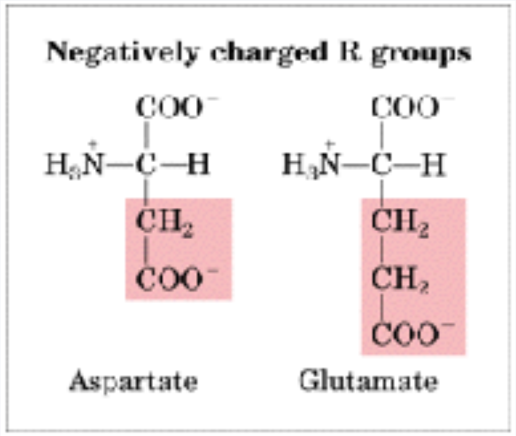
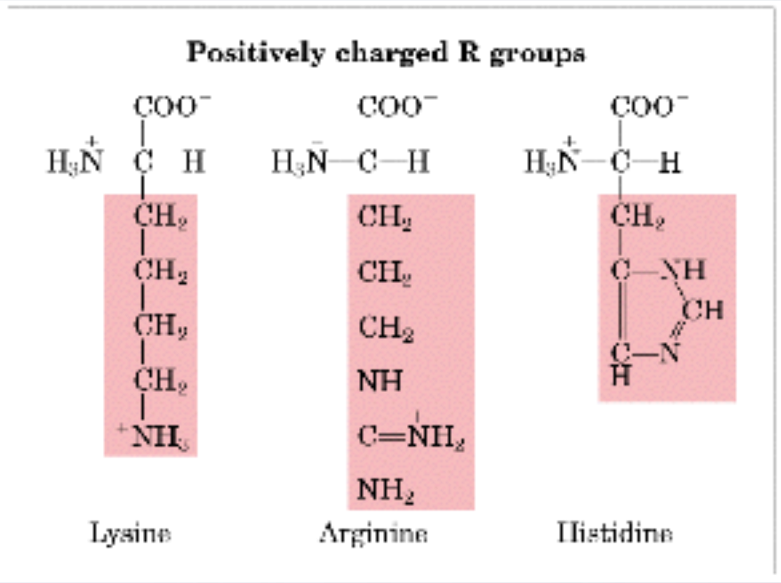


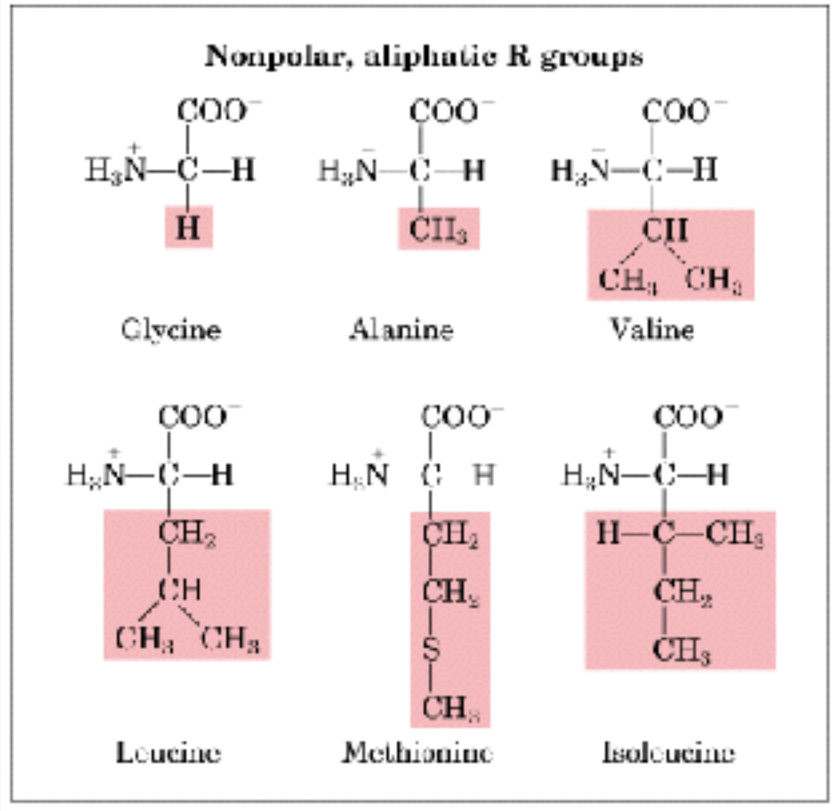
Image from: http://www.ncbi.nlm.nih.gov/books/NBK21581/

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









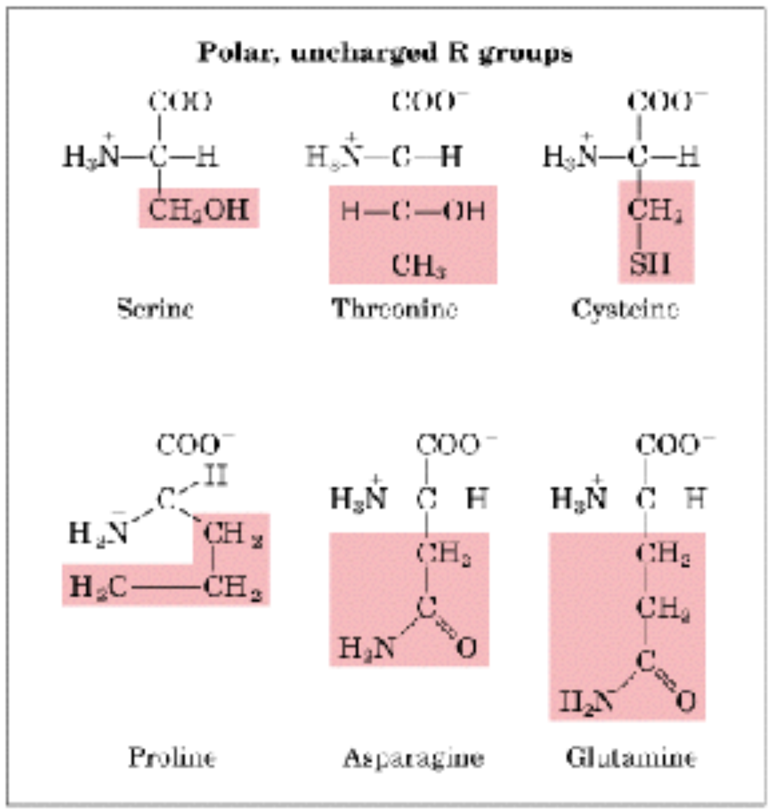
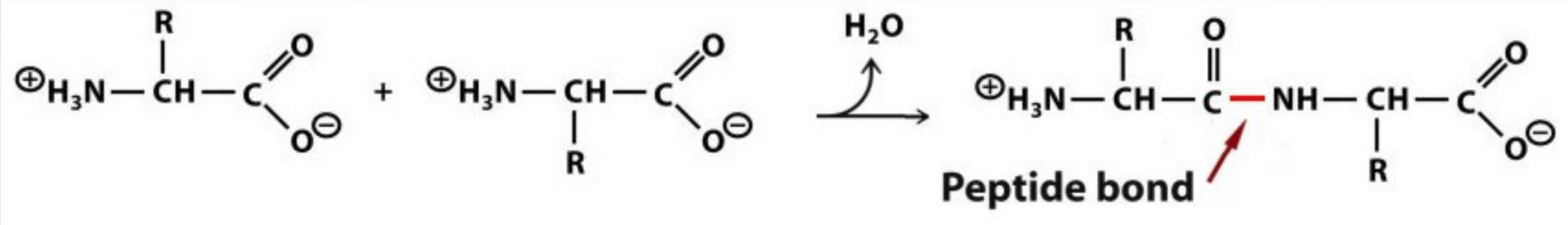


Image from: http://www.ncbi.nlm.nih.gov/books/NBK21581/

AMINO ACIDS POLYMERIZE THROUGH **PEPTIDE BOND** FORMATION



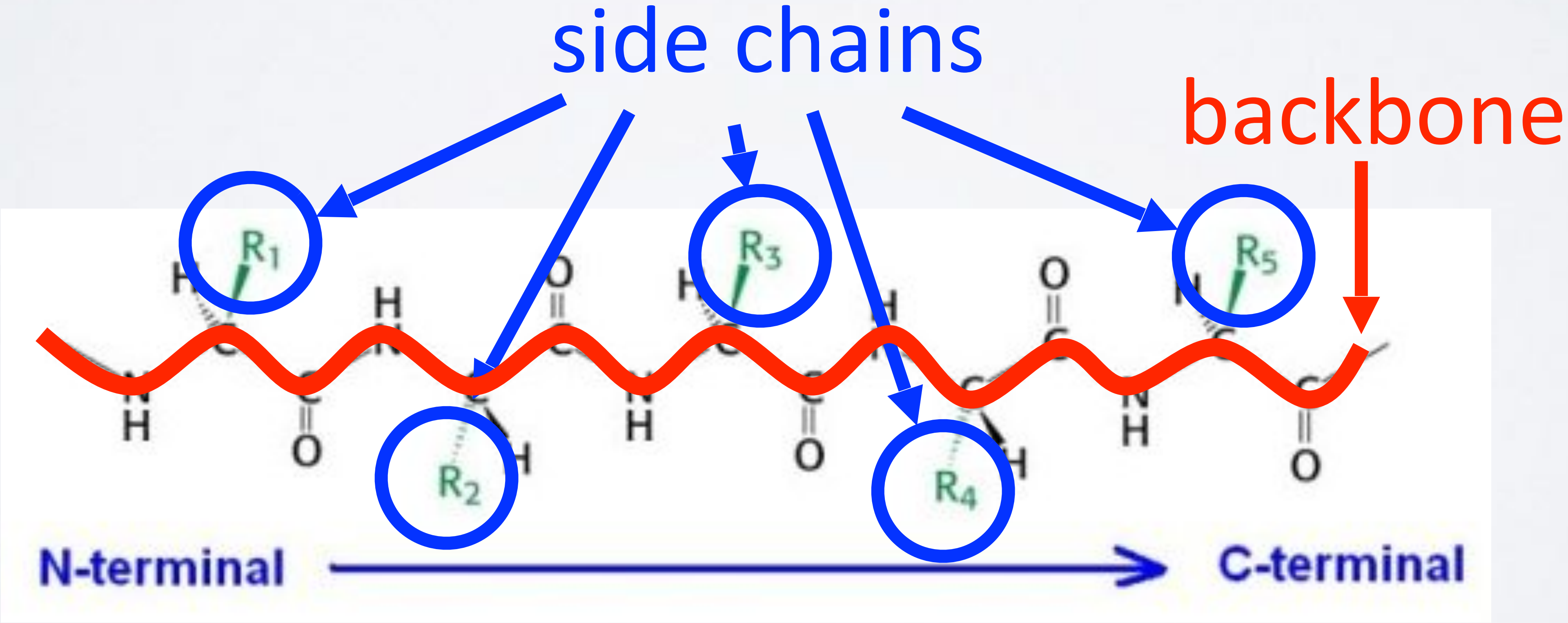
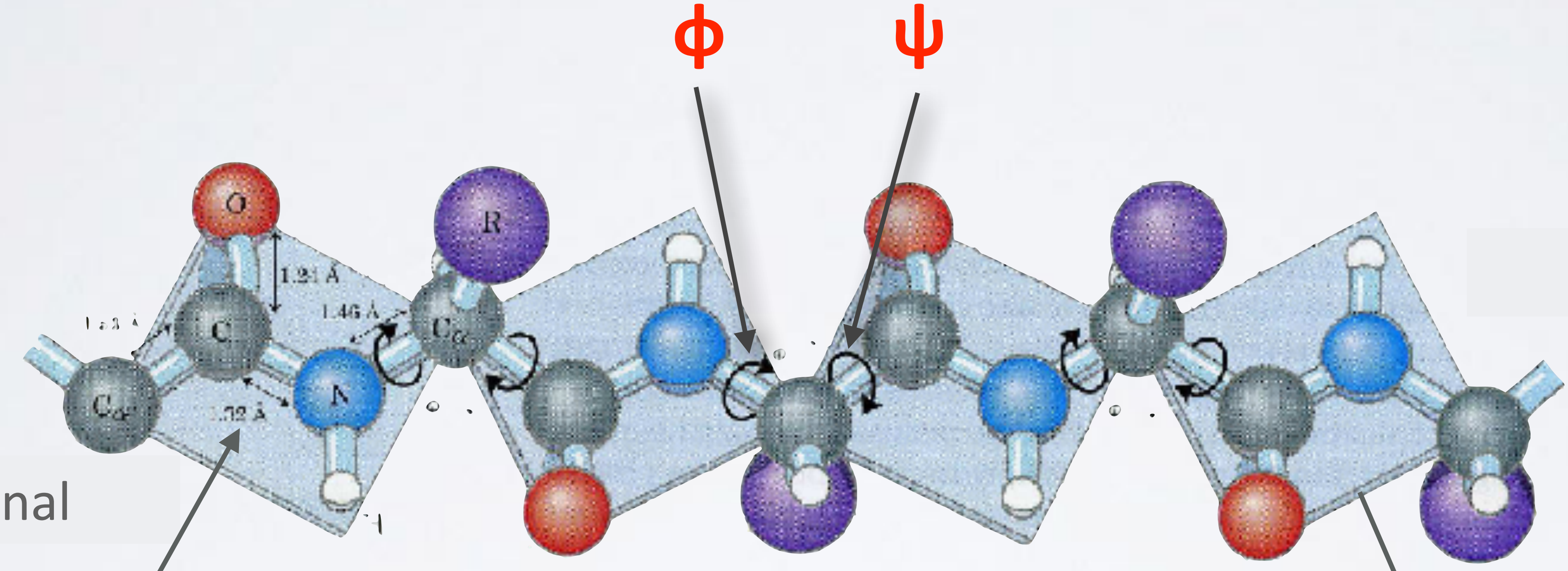


Image from: http://www.ncbi.nlm.nih.gov/books/NBK21581/

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



C-terminal

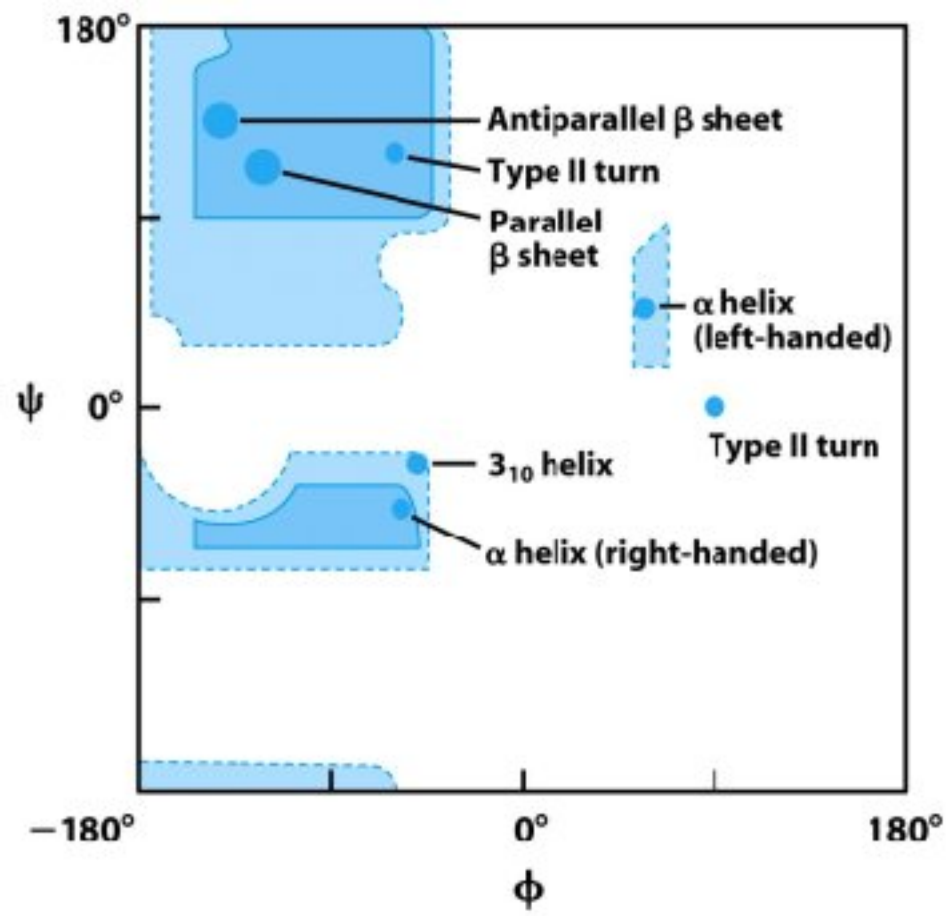
N-terminal

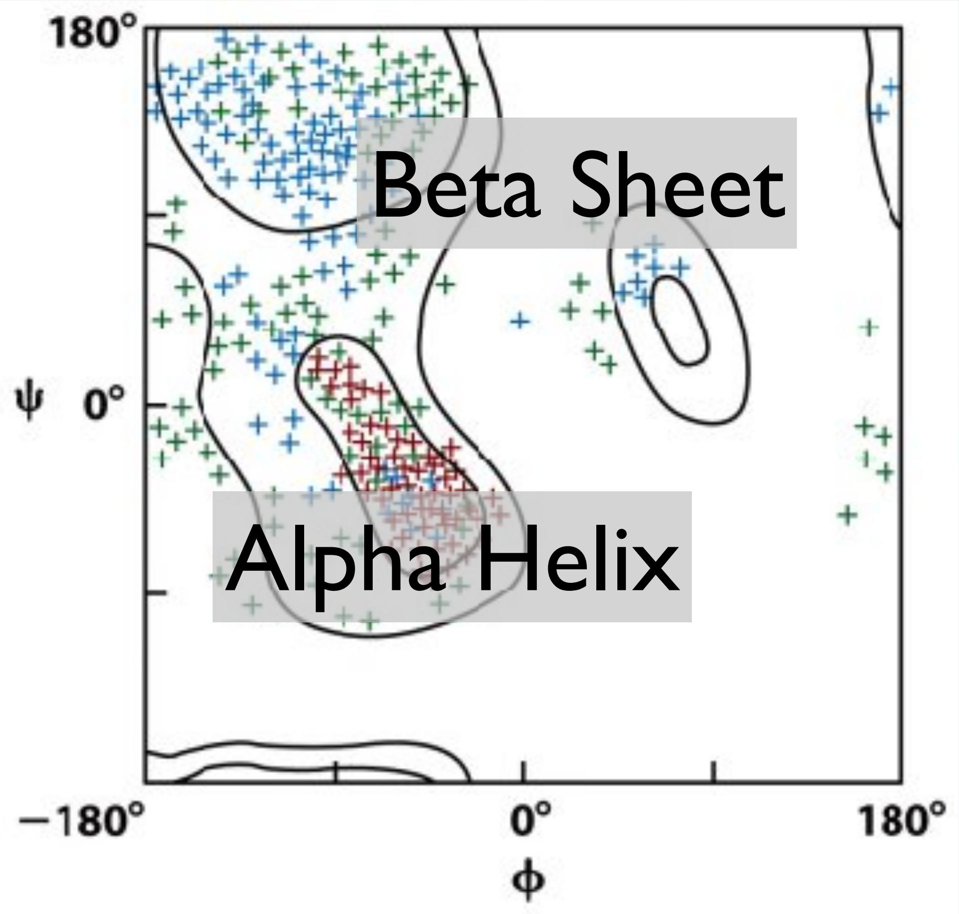
Bond angles and lengths are largely invariant

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

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

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

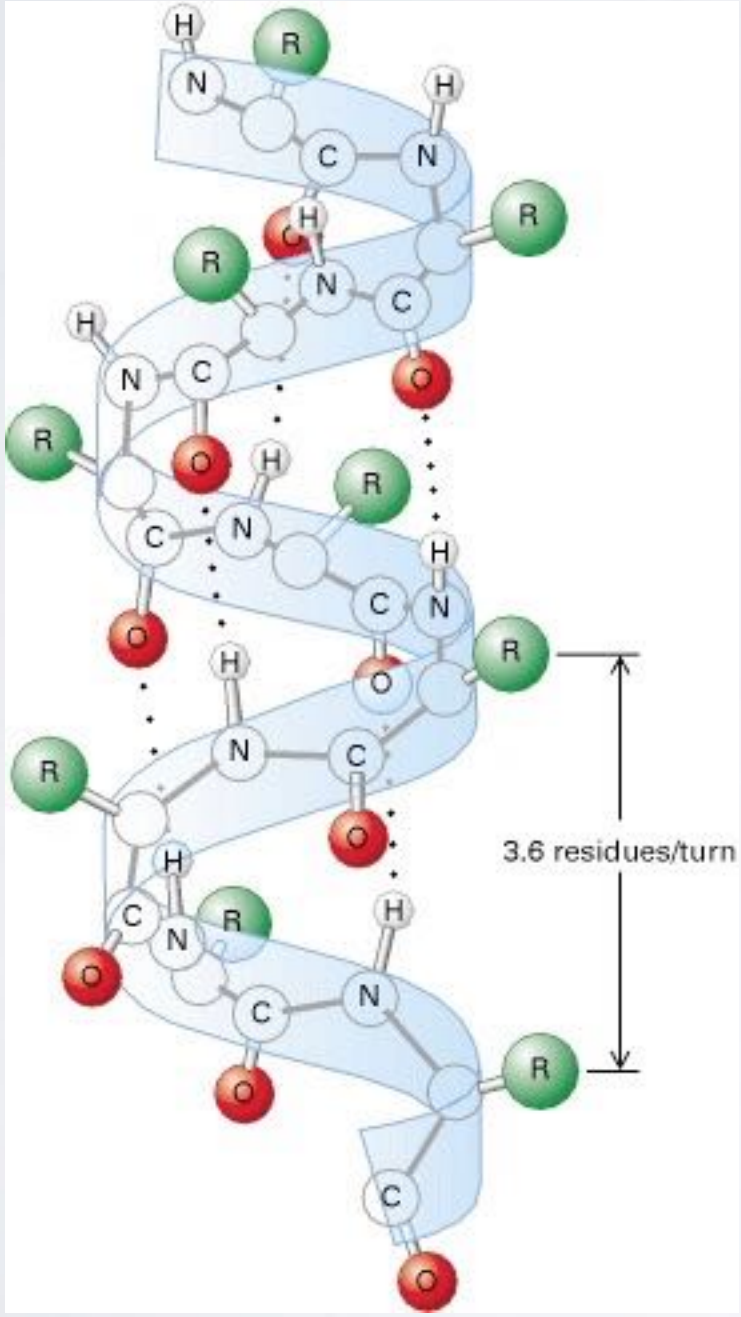




- Steric hindrance dictates torsion angle preference
- Ramachandran plot show preferred regions of φ and ψ dihedral angles which correspond to major forms of **secondary structure**

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

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



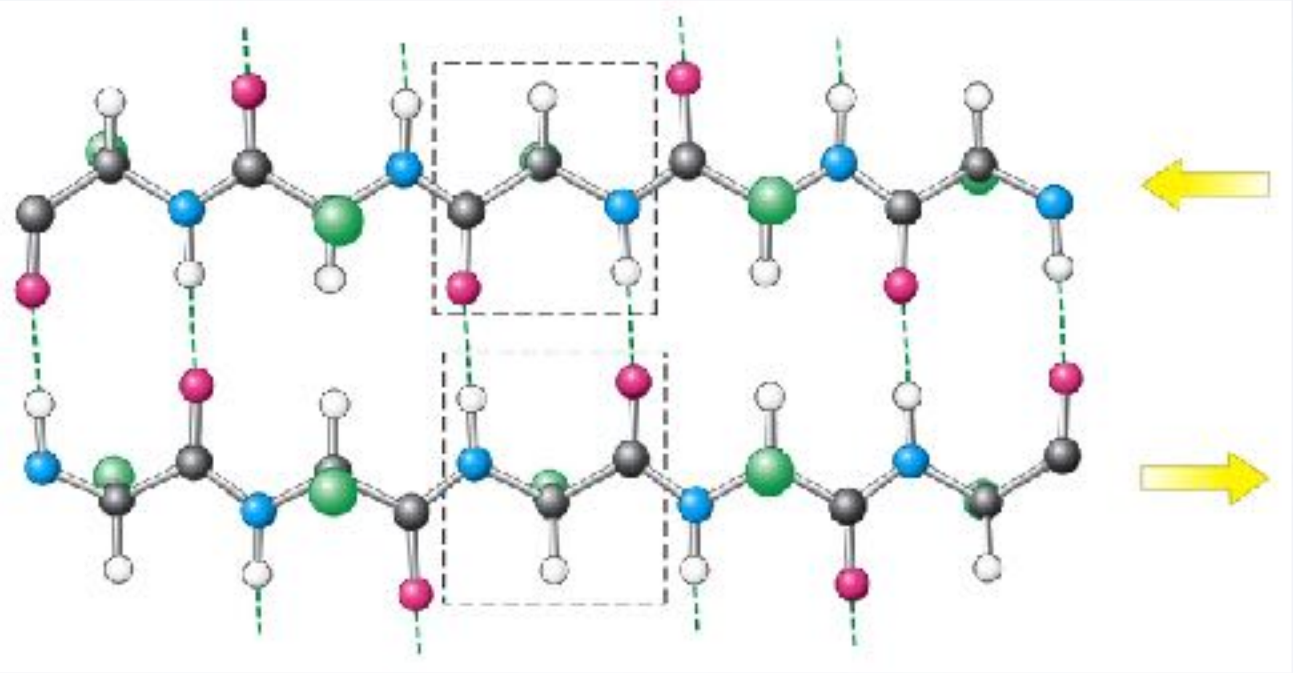
α-helix

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

Hydrogen bond: **i→i+4**

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

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

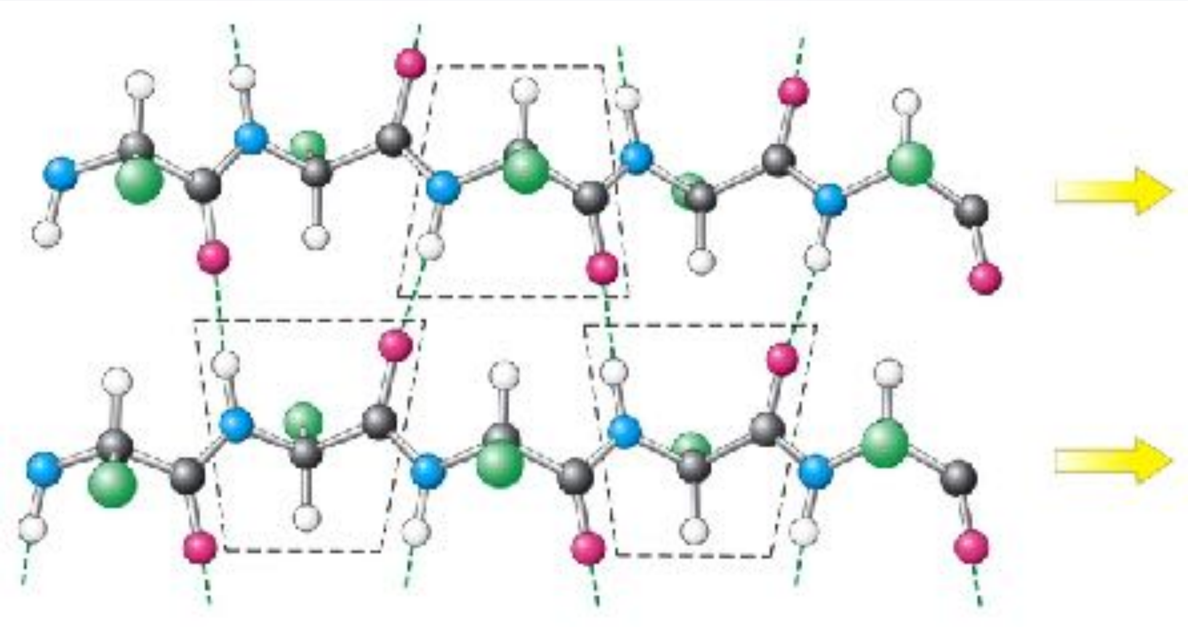


In **antiparallel** **β-sheets**

- Adjacent β-strands run in 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**

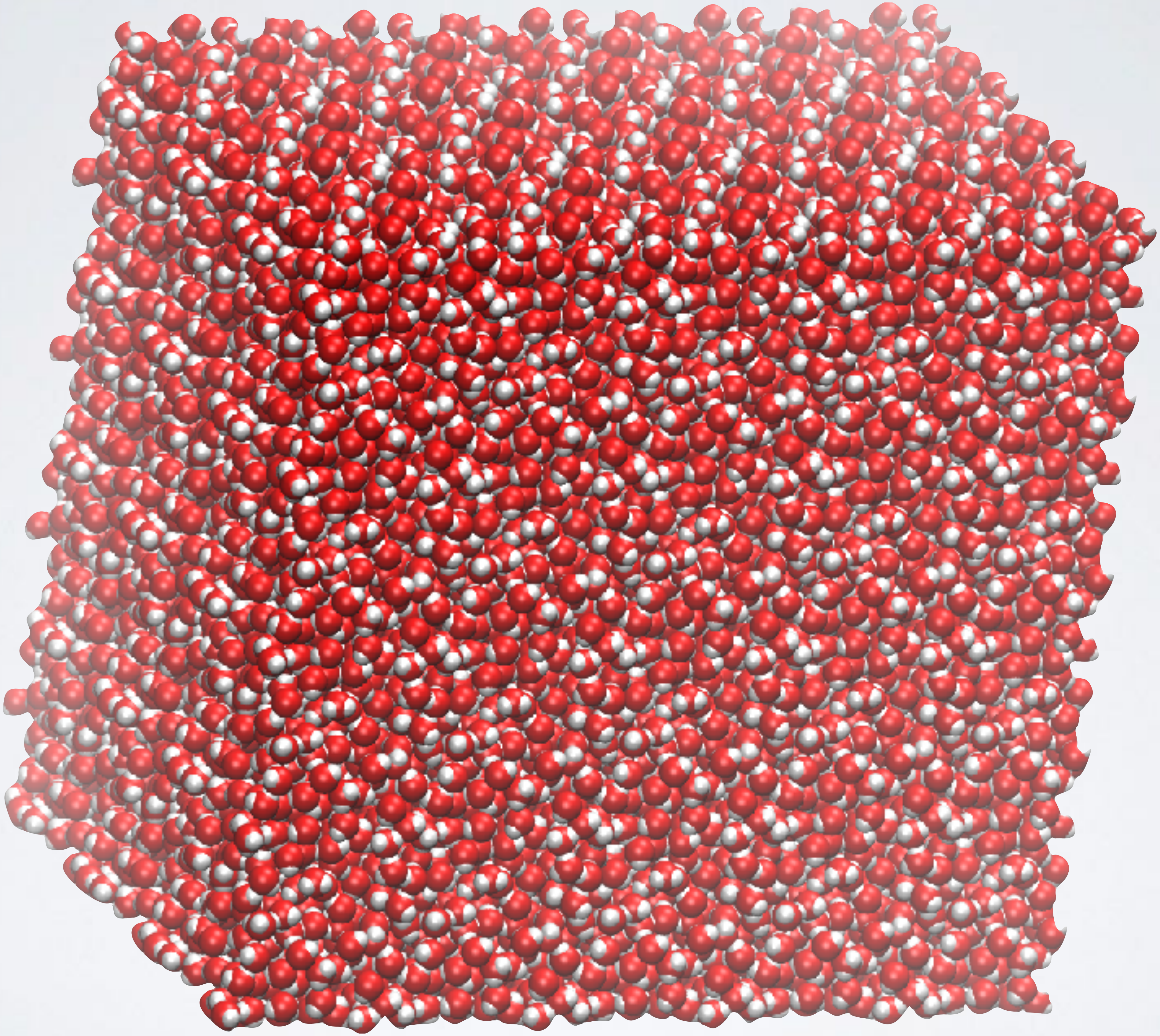


In **<u>parallel</u> β-sheets**

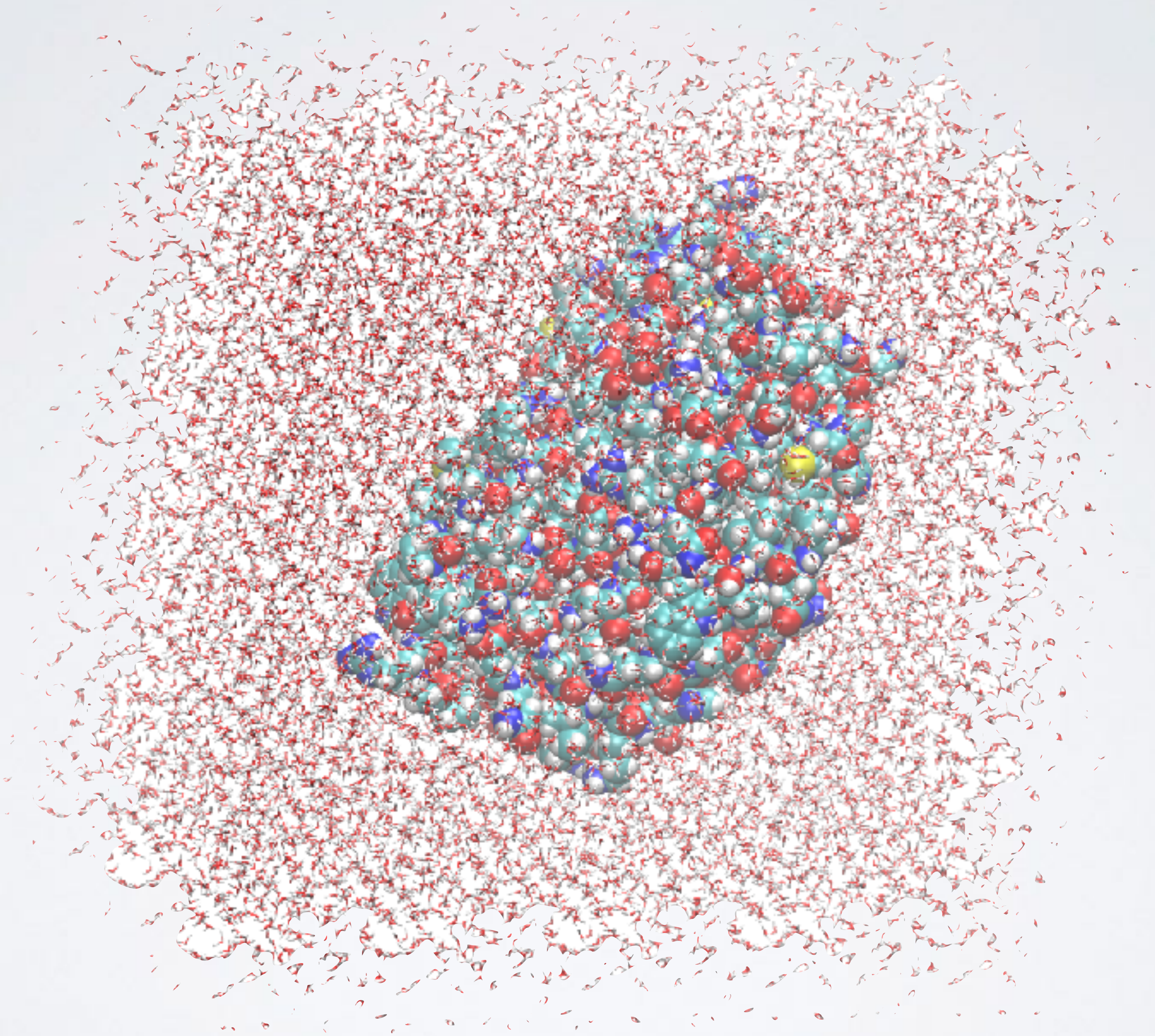
- Adjacent β-strands run in <u>same</u> direction
- Hydrogen bonds (dashed lines) between NH and CO stabilize the structure
- The side chains (in green) are above and below the sheet

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

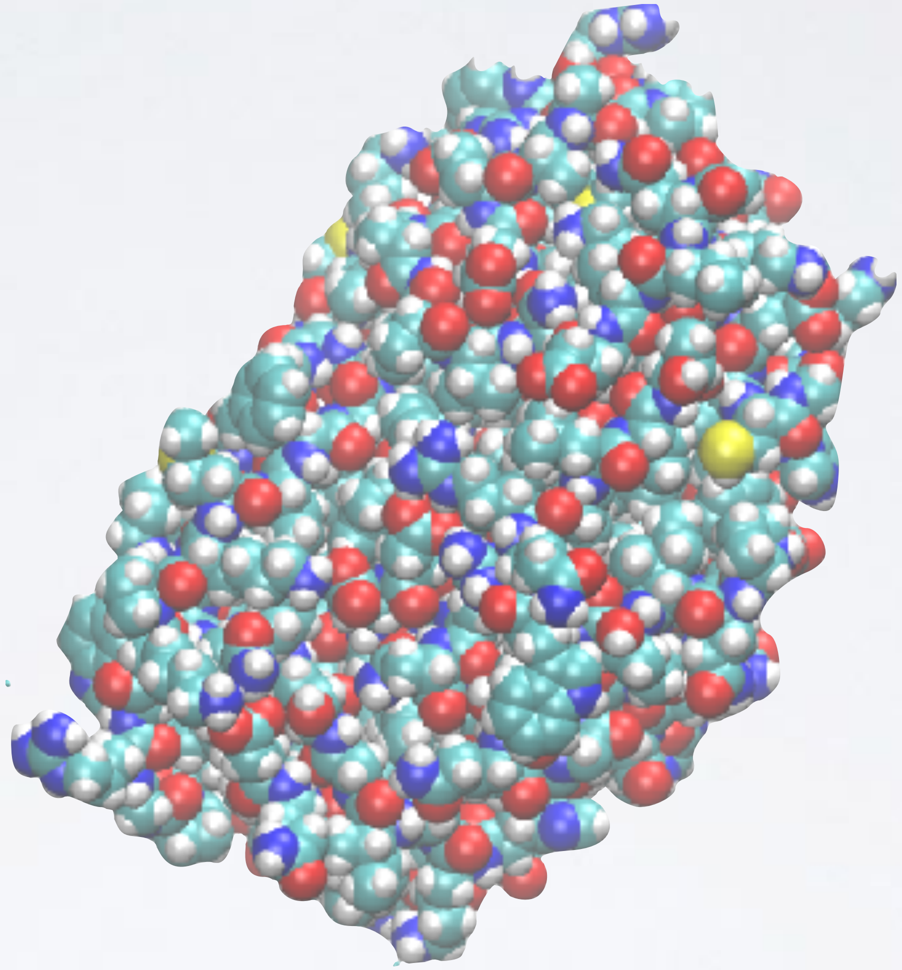
What Does a Protein Look like?



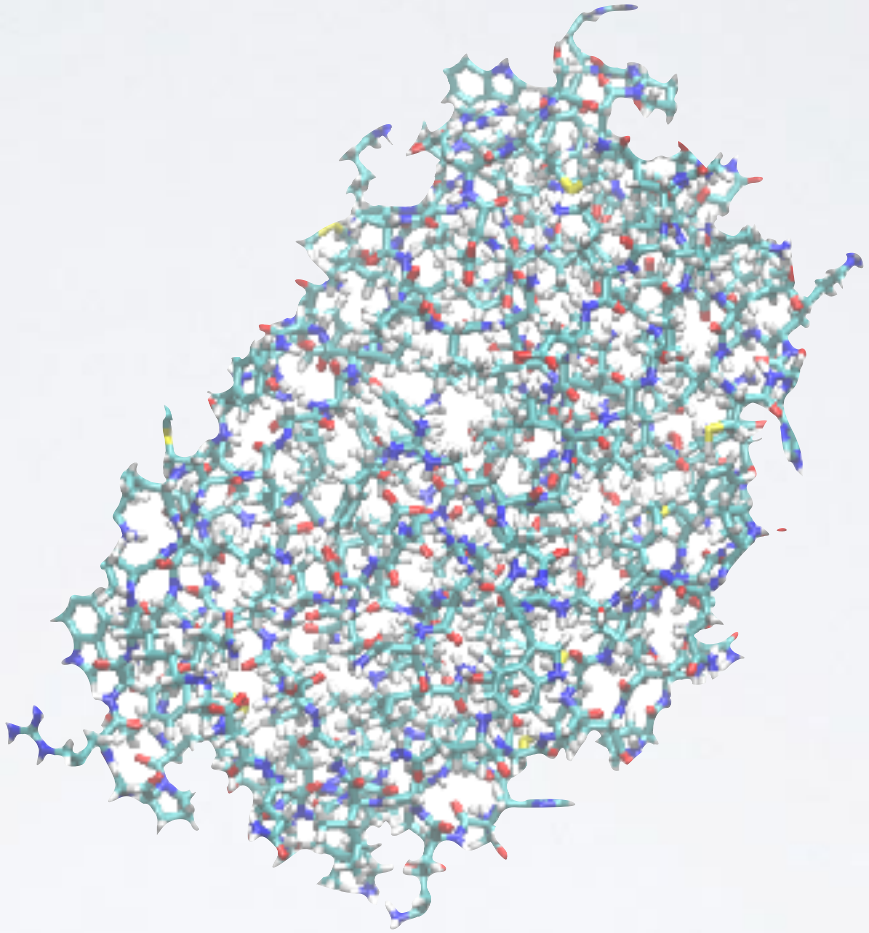
- Proteins are stable (and hidden) in water



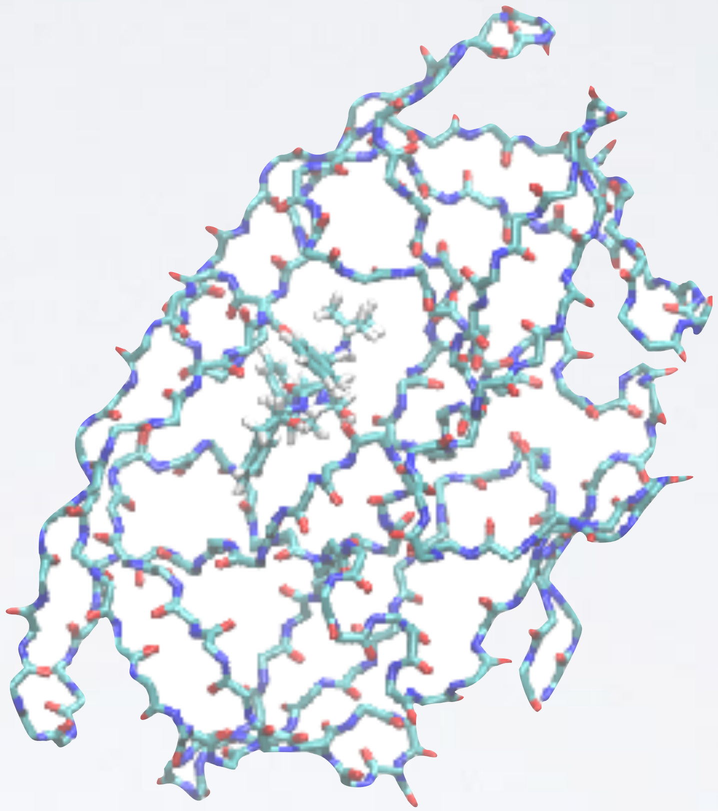
- Proteins closely interact with water



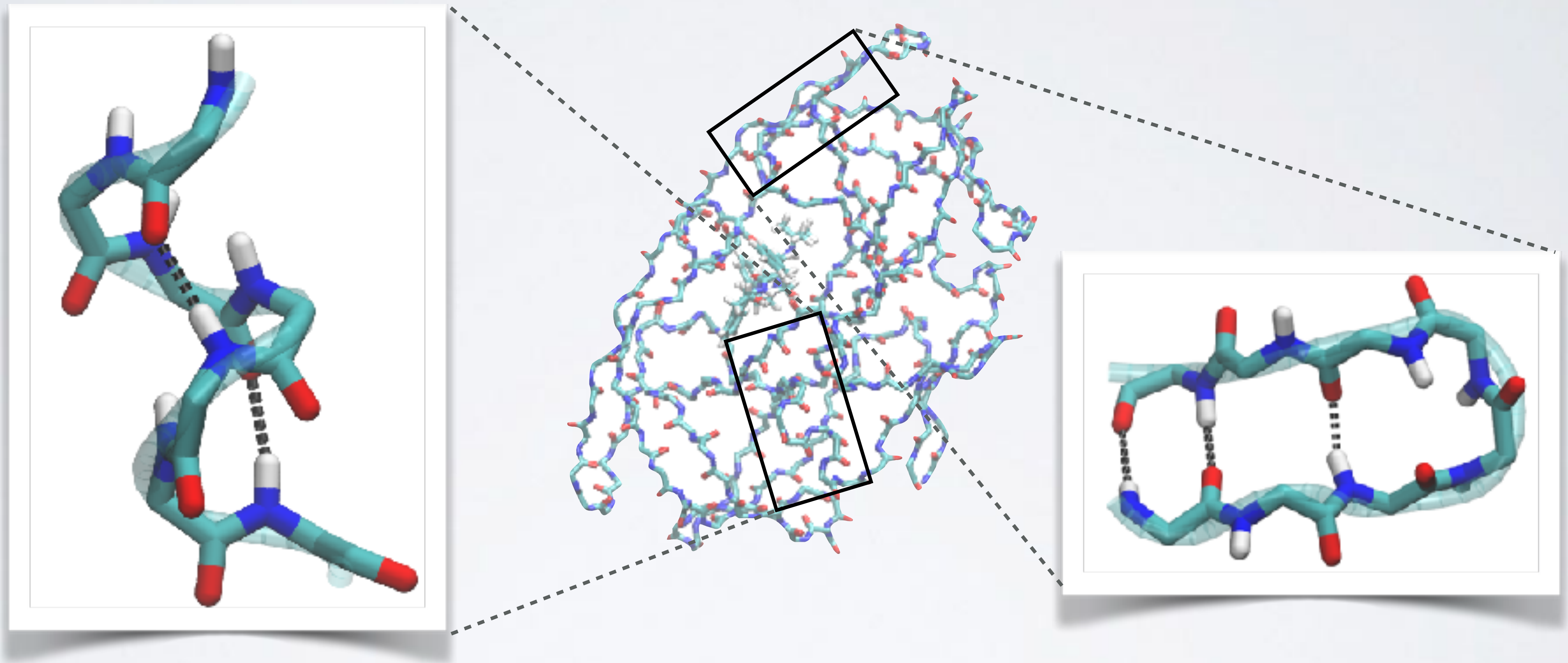
- Proteins are close packed solid but flexible objects (globular)



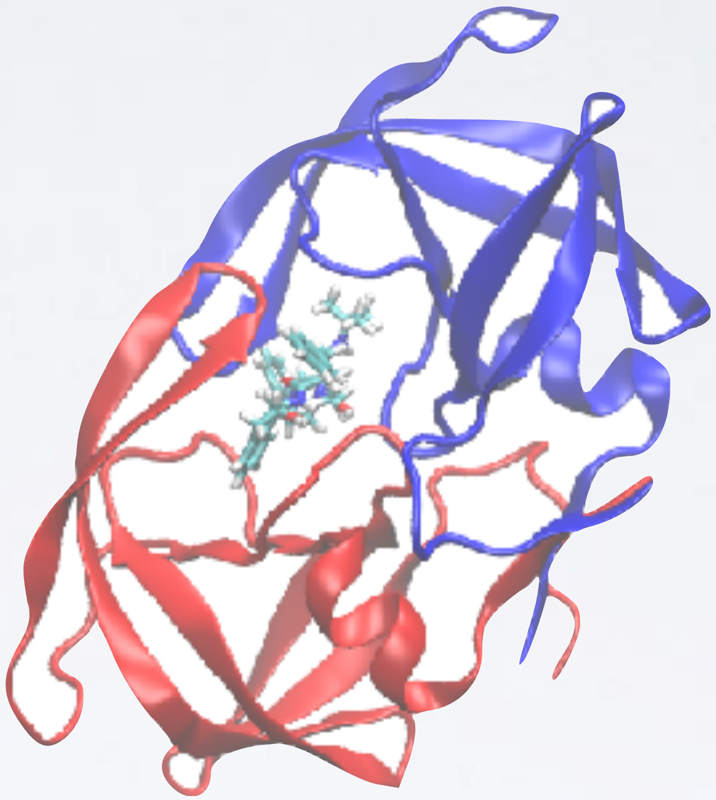
- Due to their large size and complexity it is often hard to see whats important in the structure



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

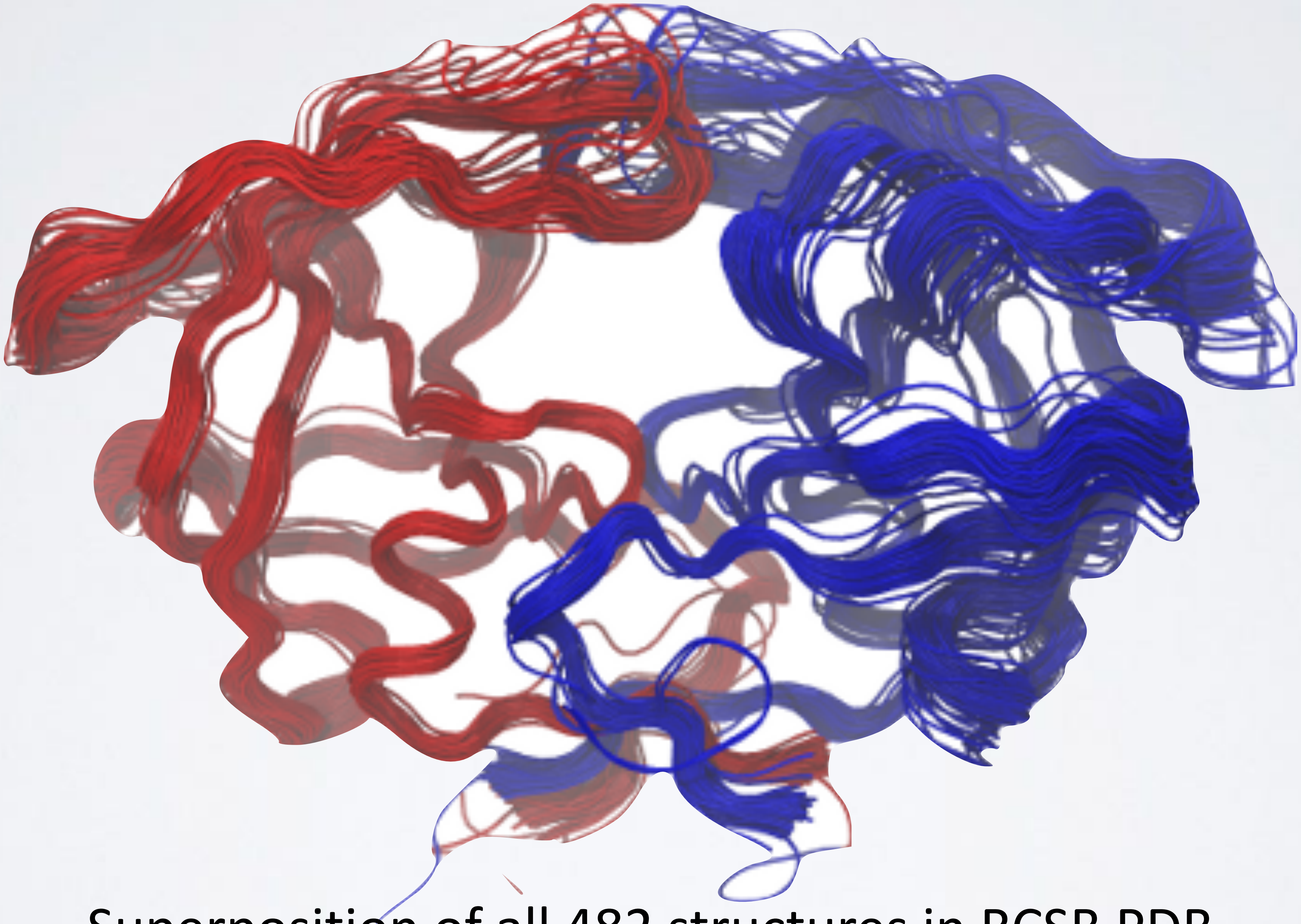


- Backbone or main-chain representation can help trace chain topology & reveal secondary structure



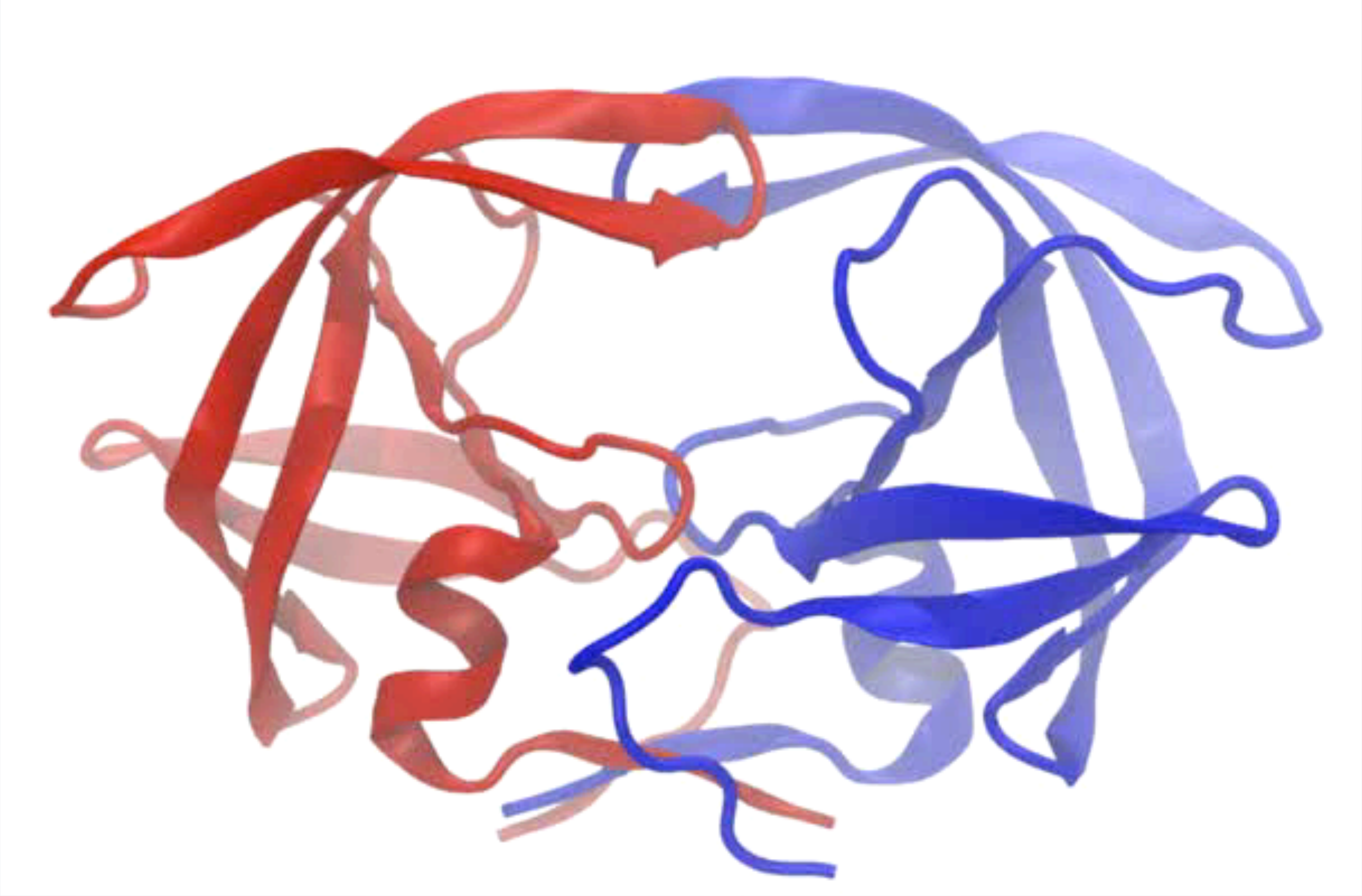
- Simplified secondary structure representations are commonly used to communicate structural details
- Now we can clearly see 2^o, 3^o and 4^o structure
- Coiled chain of connected secondary structures

DISPLACEMENTS REFLECT INTRINSIC FLEXIBILITY



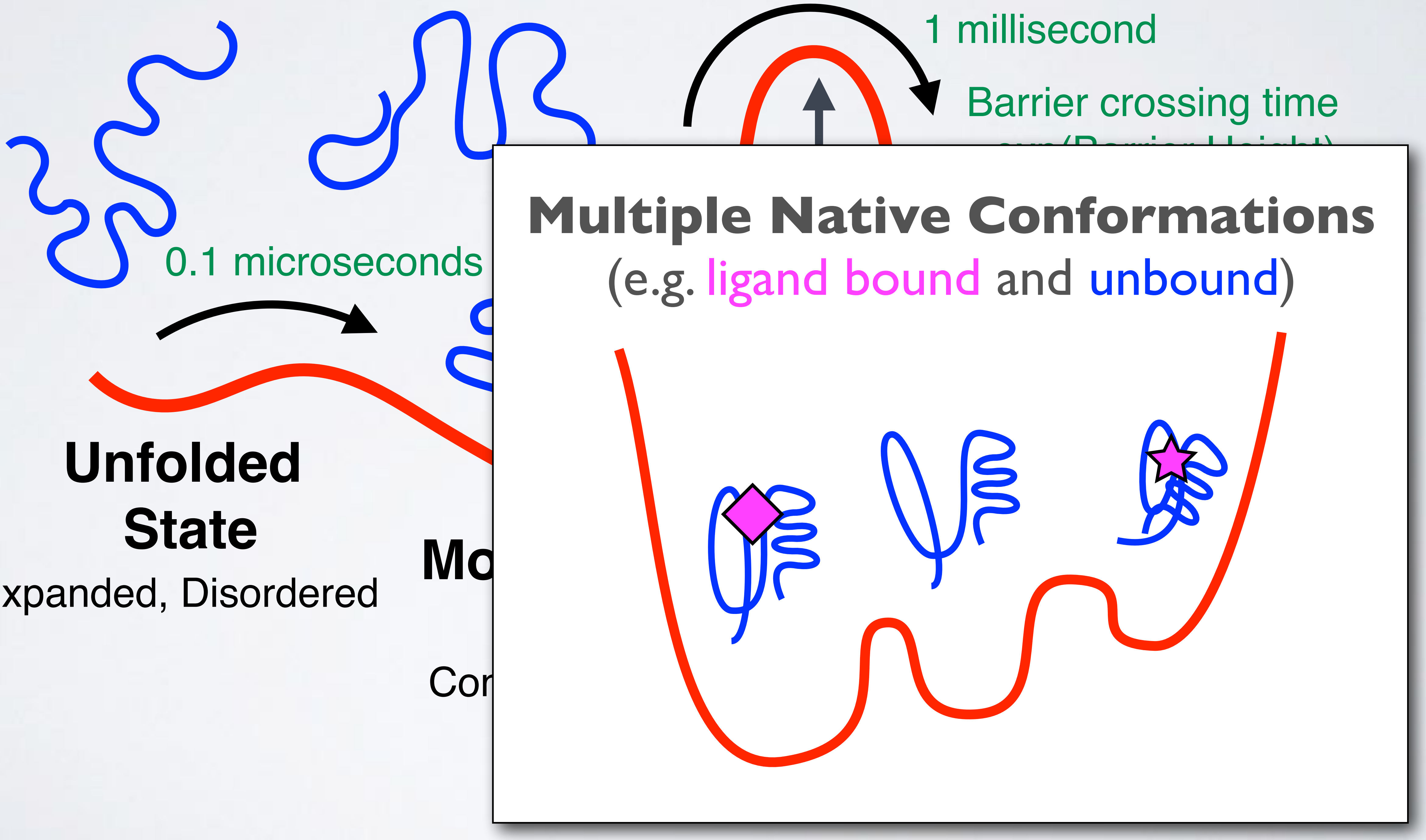
Superposition of all 482 structures in RCSB PDB
(23/09/2015)

DISPLACEMENTS REFLECT INTRINSIC FLEXIBILITY



Principal component analysis (PCA) of experimental structures

KEY CONCEPT: **ENERGY LANDSCAPE**

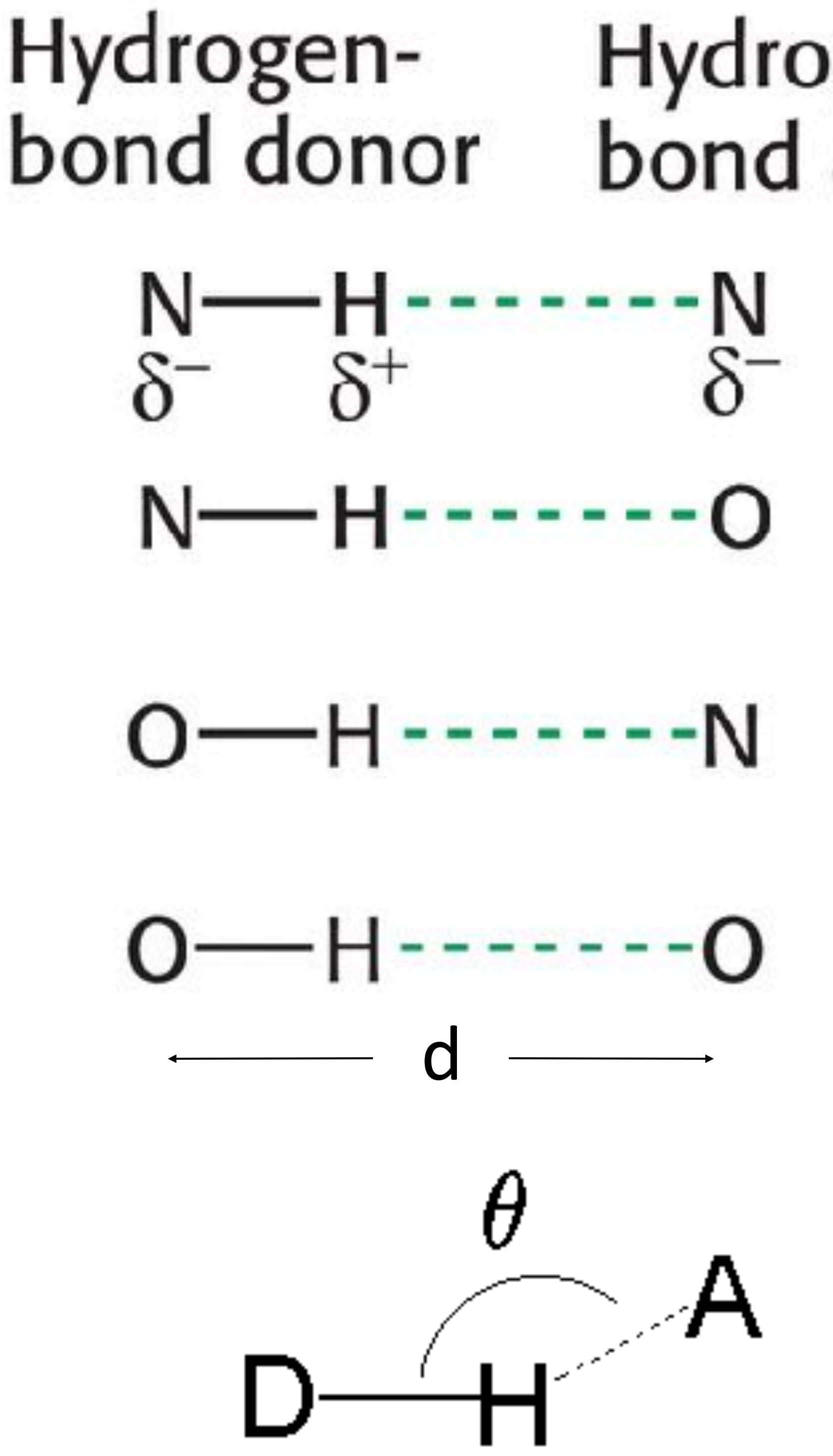


Multiple Native Conformations

(e.g. ligand bound and unbound)

Key forces affecting structure:

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

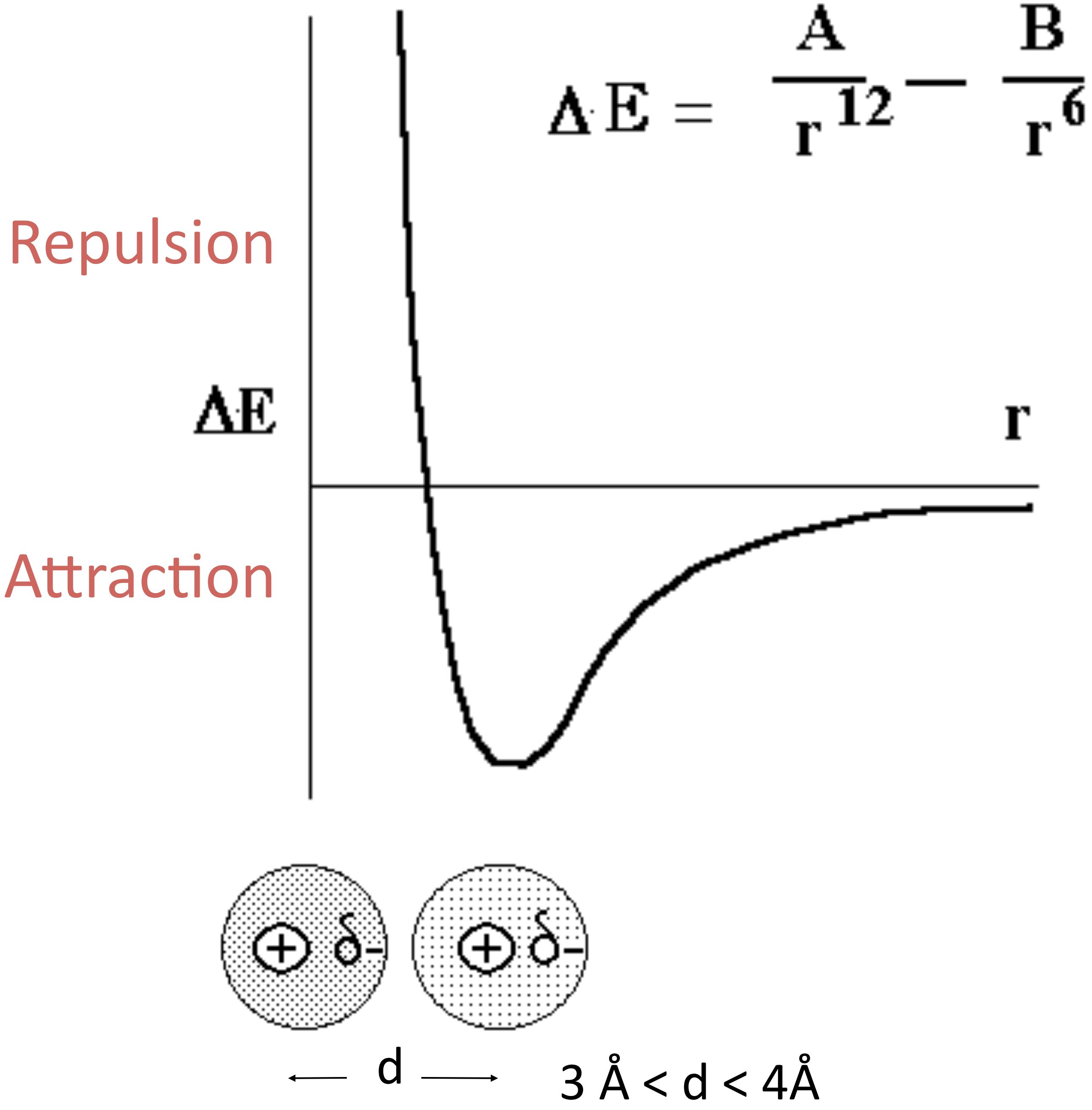


2.6 Å < d < 3.1Å

150° < θ < 180°

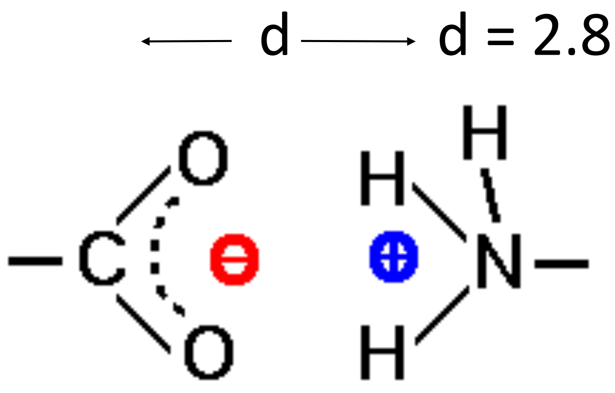
Key forces affecting structure:

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



Key forces affecting structure:

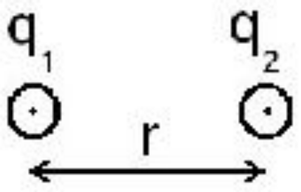
- H-bonding
- Van der Waals
- Electrostatics
- Hydrophobicity



d = 2.8 Å

carboxyl group and amino group

(some time called IONIC BONDs or SALT BRIDGEs)



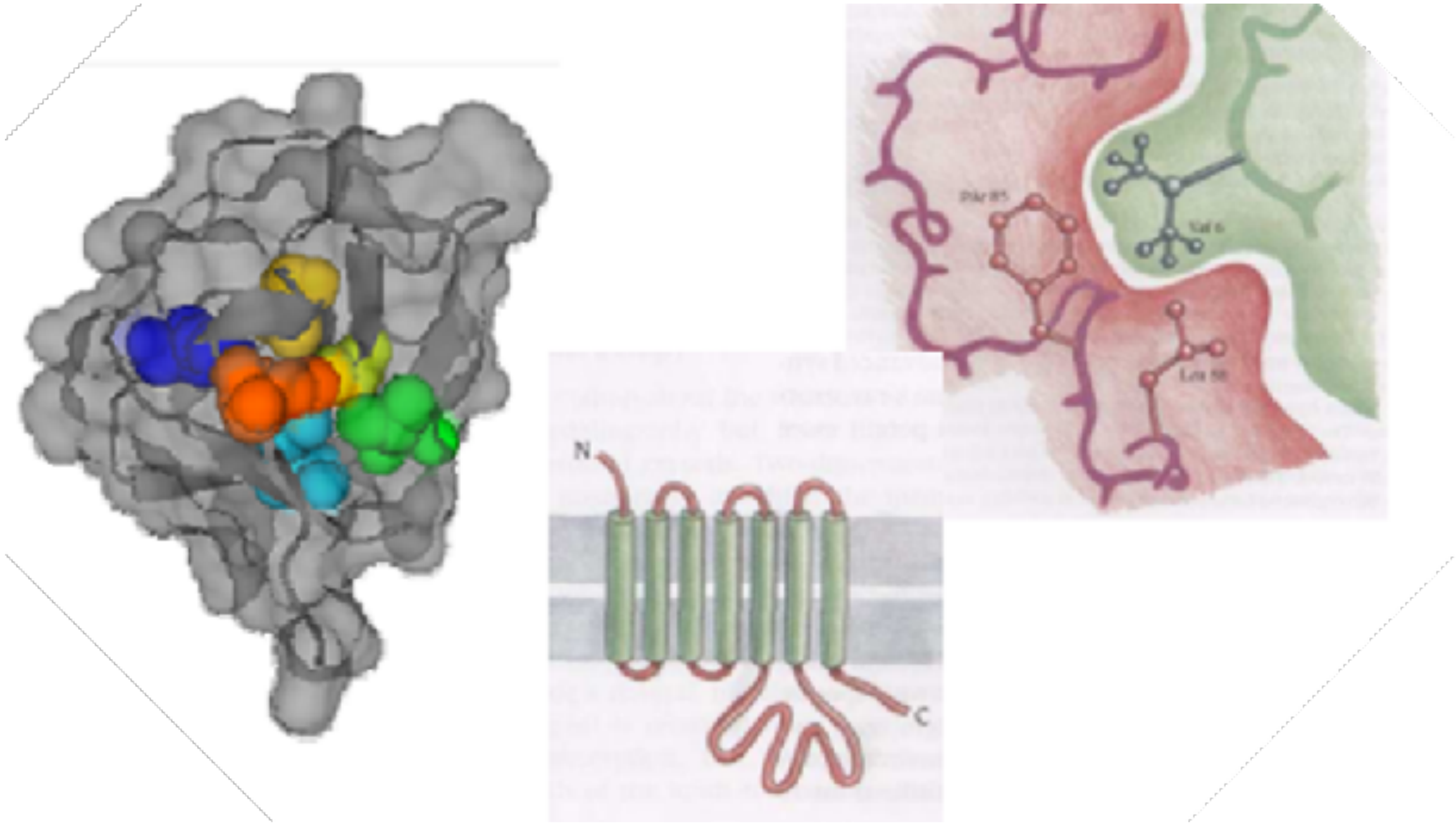
Coulomb's law

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

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



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.

NEXT UP:

- **Overview of structural bioinformatics**
    - Major motivations, goals and challenges
- **Fundamentals of protein structure**
    - Composition, form, forces and dynamics
- **Representing and interpreting protein structure**
    - Modeling energy as a function of structure
- **Example application areas**
    - Predicting functional dynamics & drug discovery

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

Two main approaches:

(1). **Physics-Based**

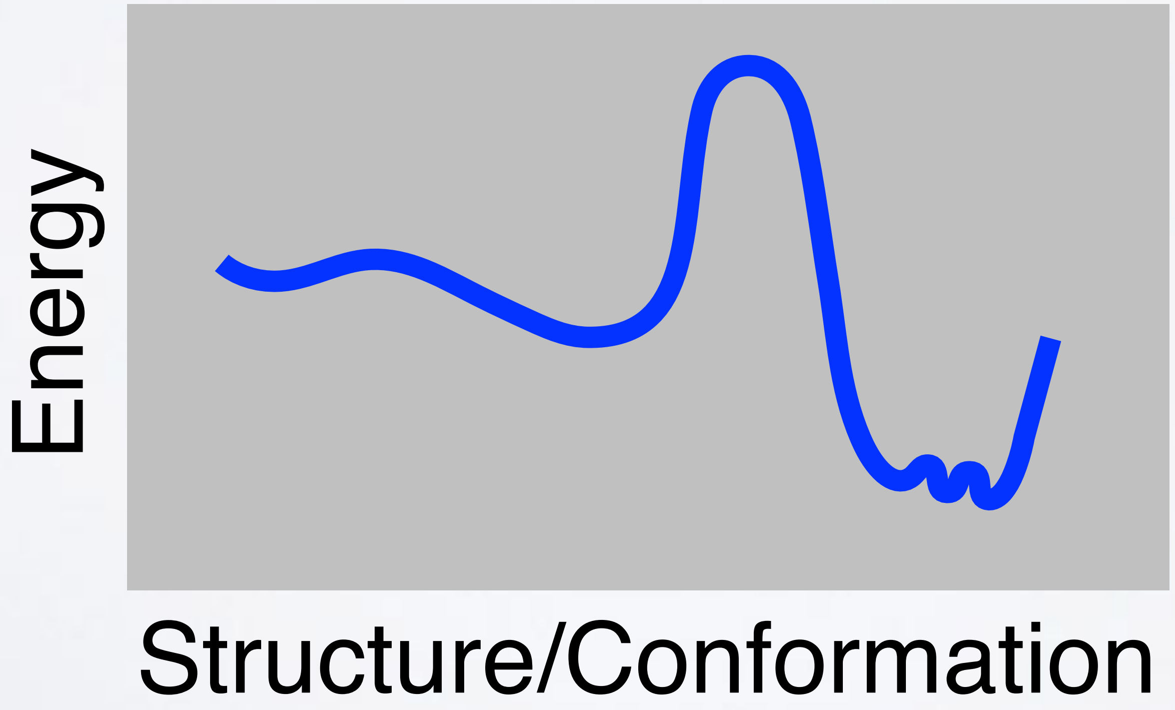
(2). **Knowledge-Based**

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

Two main approaches:

(1). **Physics-Based**

(2). **Knowledge-Based**

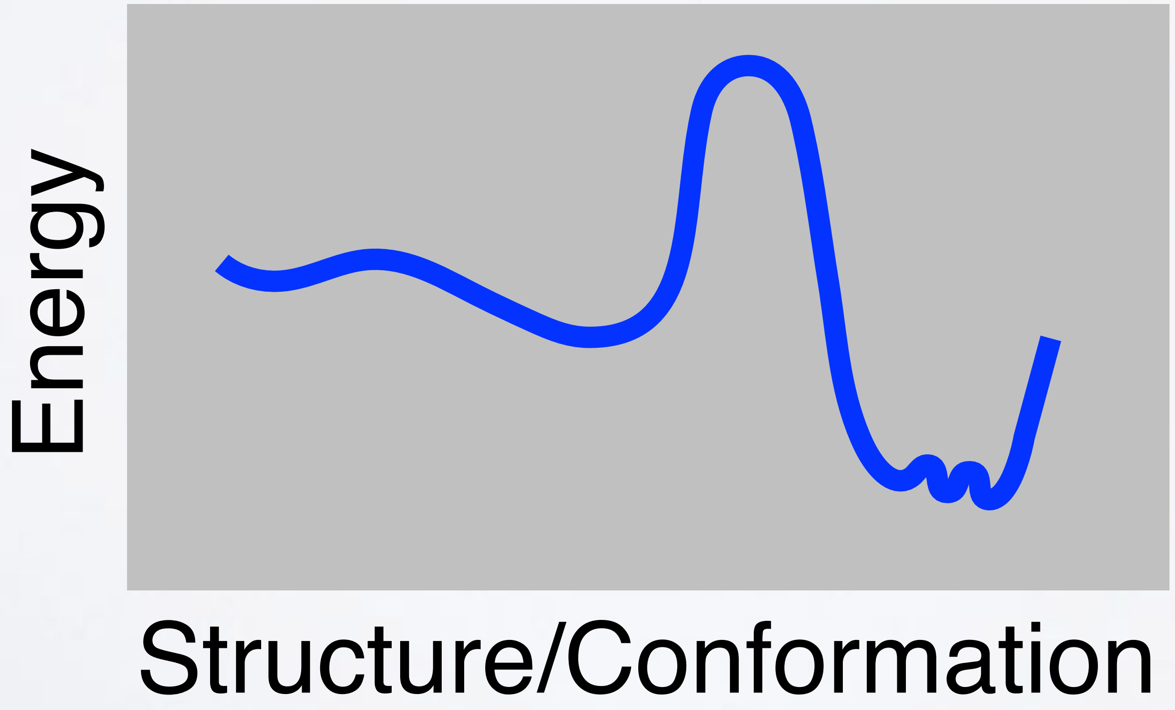


KEY CONCEPT: POTENTIAL FUNCTIONS DESCRIBE A SYSTEMS **ENERGY** AS A FUNCTION OF ITS **STRUCTURE**

Two main approaches:

(1). **Physics-Based**

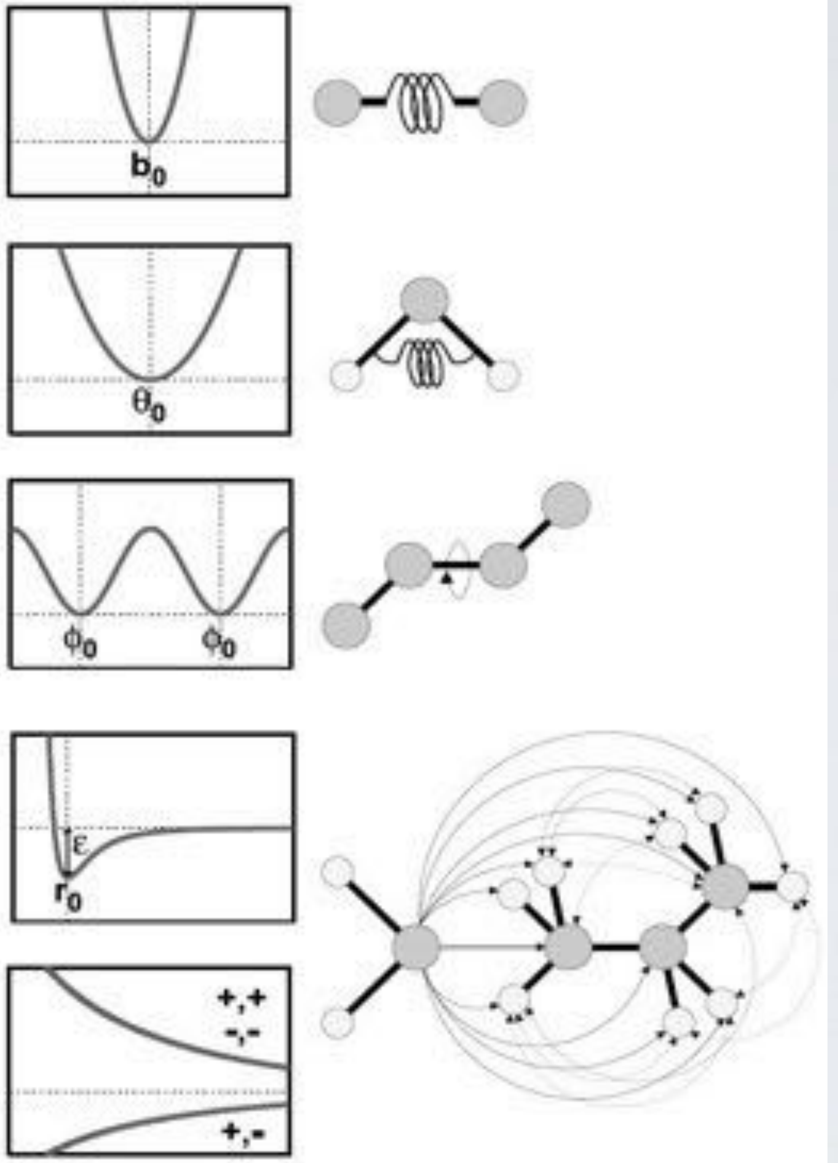
(2). **Knowledge-Based**



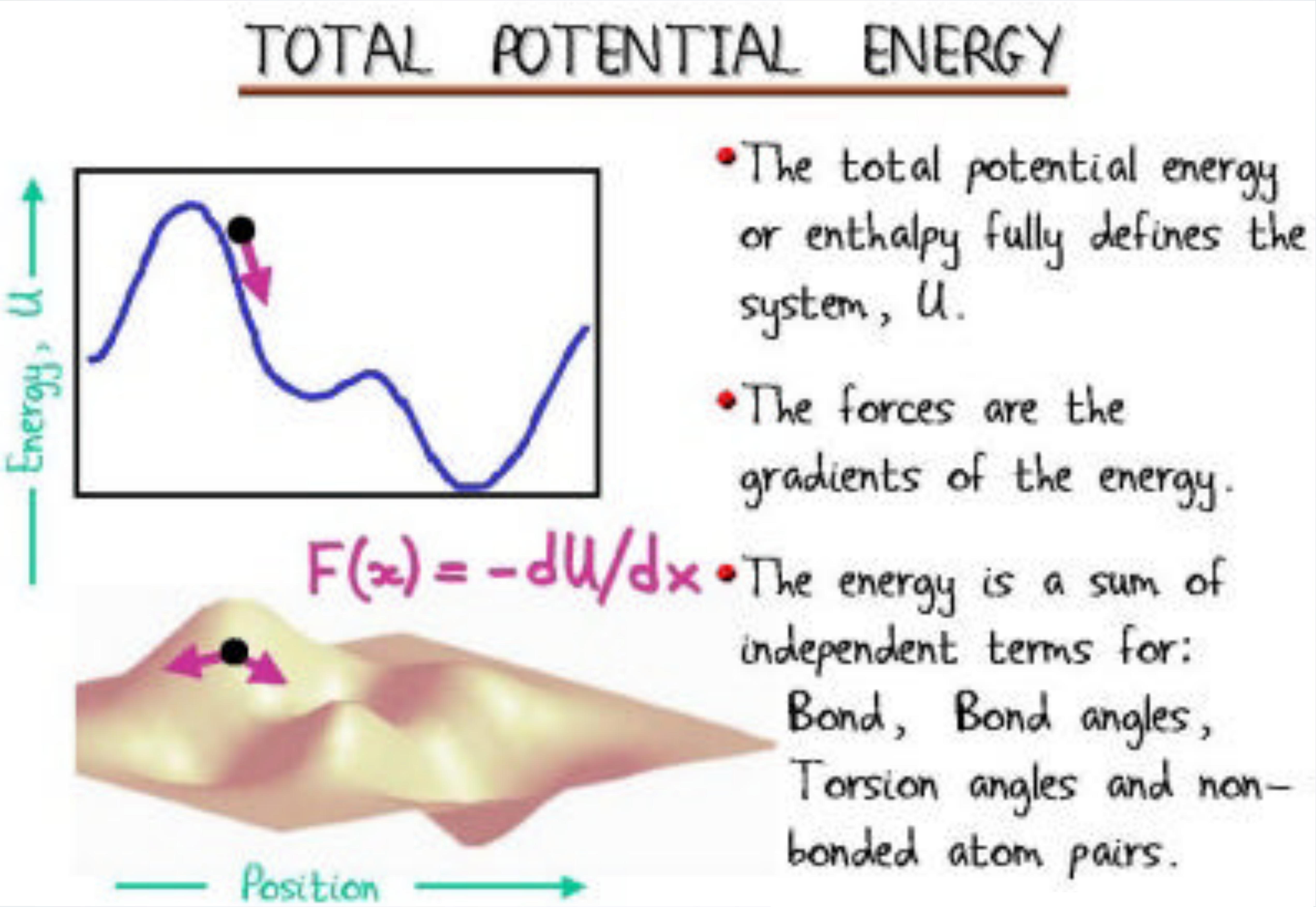
PHYSICS-BASED POTENTIALS
ENERGY TERMS FROM PHYSICAL THEORY

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

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/



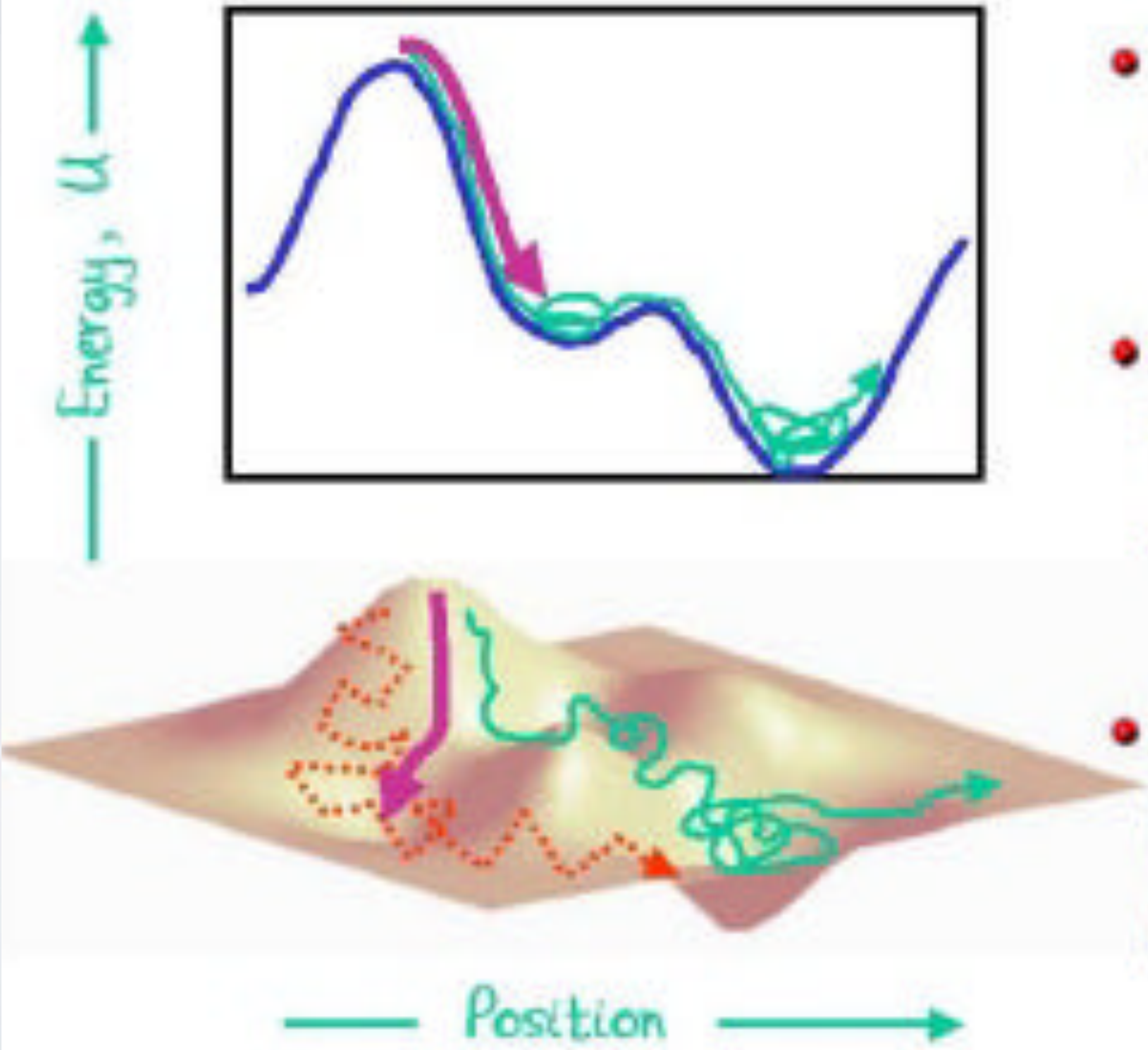
TOTAL POTENTIAL ENERGY

- The total potential energy or enthalpy fully defines the system, U.
- The forces are the gradients of the energy.
- The energy is a sum of independent terms for: Bond, Bond angles, Torsion angles and non-bonded atom pairs.

$F(x) = -dU/dx$

Slide Credit: Michael Levitt

MOVING OVER THE ENERGY SURFACE



- Energy Minimization drops into local minimum.
- Molecular Dynamics uses thermal energy to move smoothly over surface.
- Monte Carlo Moves are random. Accept with probability $\exp(-\Delta U/kT)$.

Slide Credit: Michael Levitt

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

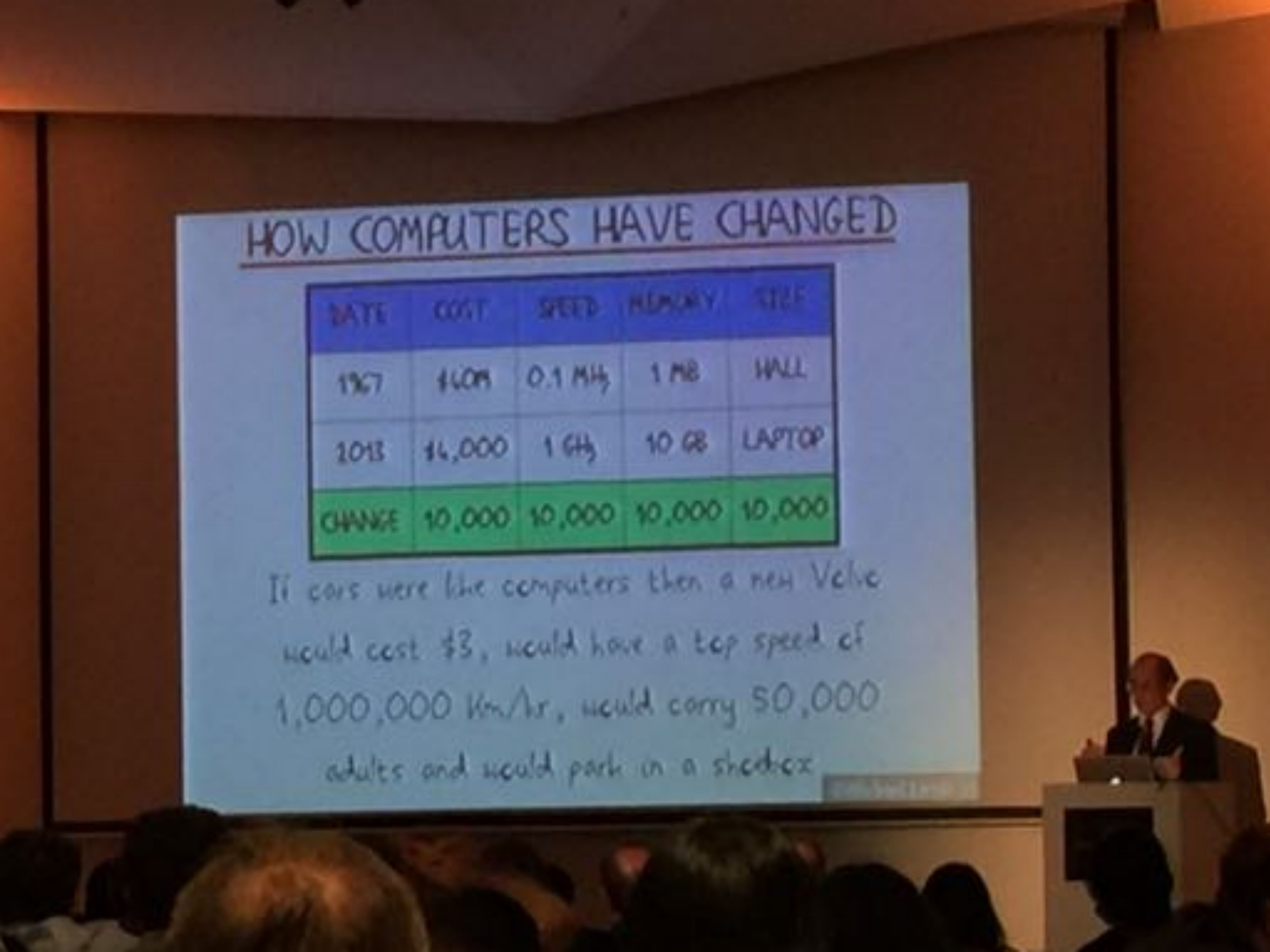
Status

Useful, widely adopted but far from perfect

Multiple groups working on fewer, better approxs

Force fields, quantum
entropy, water effects

Moore's law: hardware improving

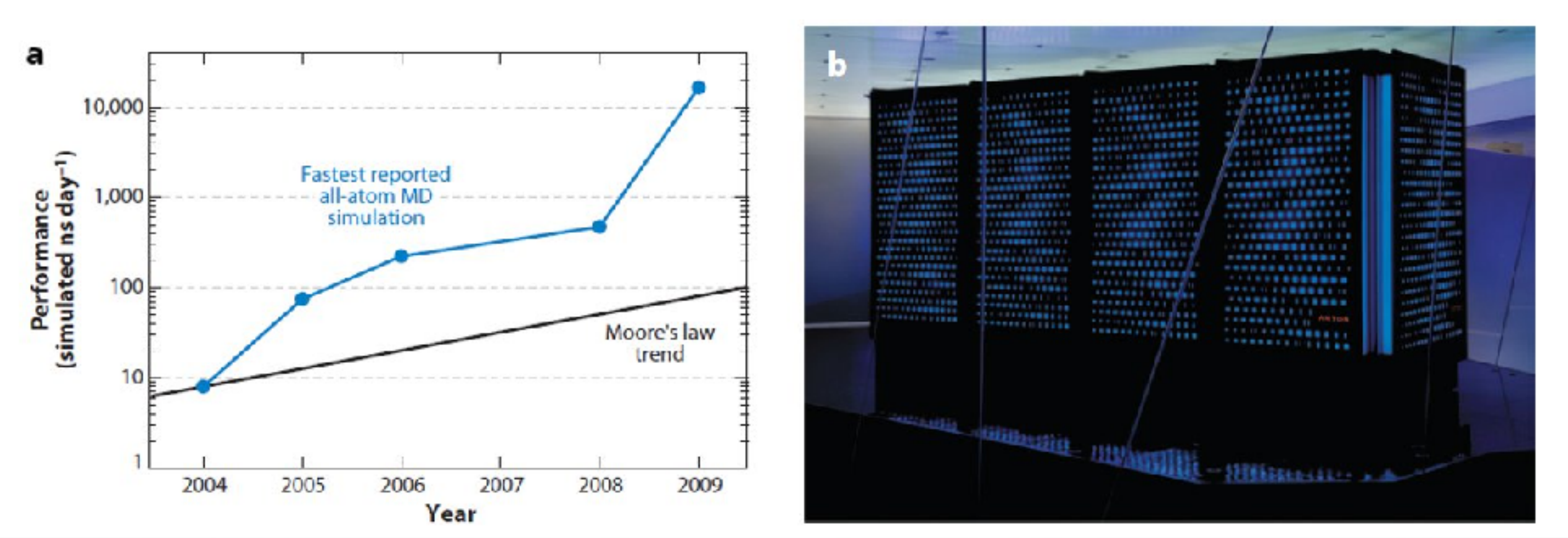


HOW COMPUTERS HAVE CHANGED

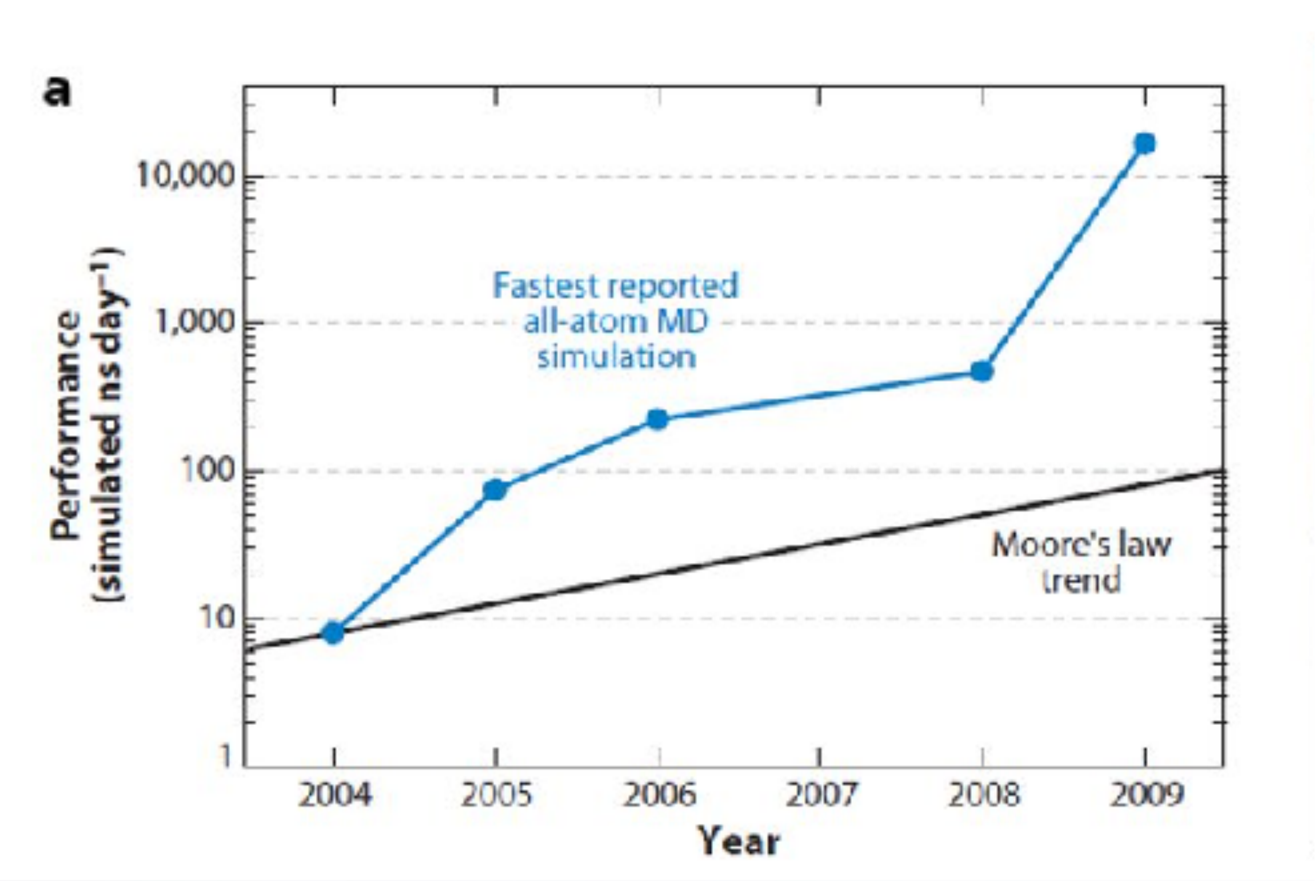
DATE	COST	SPEED	MEMORY	SIZE
1967	$60M	0.1 MHz	1 MB	HALL
2013	$6,000	1 GHz	10 GB	LAPTOP
CHANGE	10,000	10,000	10,000	10,000

If cars were like the computers then a new Volvo would cost $3, would have a top speed of 1,000,000 Km/hr, would carry 50,000 adults and would park in a shoebox

SIDE-NOTE: GPUS AND ANTON SUPERCOMPUTER



SIDE-NOTE: GPUS AND ANTON SUPERCOMPUTER



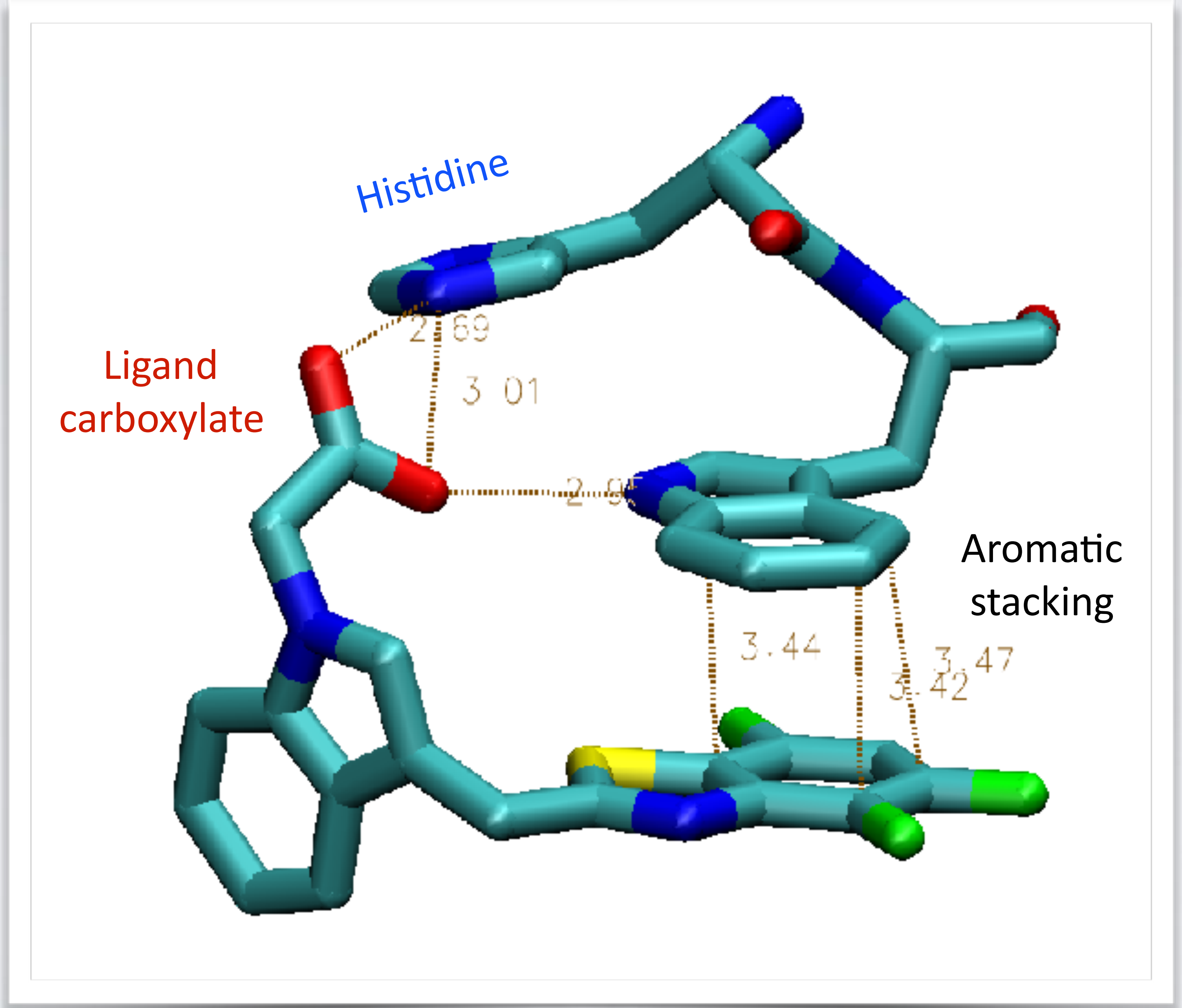
KEY CONCEPT: POTENTIAL FUNCTIONS DESCRIBE A SYSTEMS **ENERGY** AS A FUNCTION OF ITS **STRUCTURE**

Two main approaches:

(1). **Physics-Based**

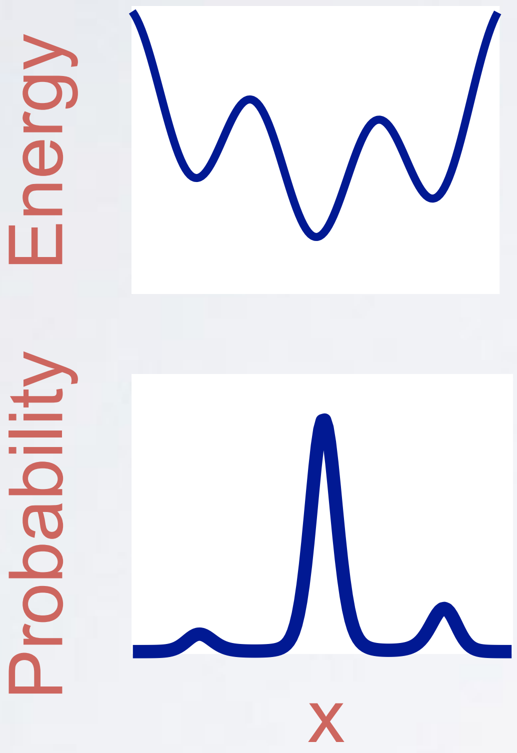
(2). **Knowledge-Based**

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

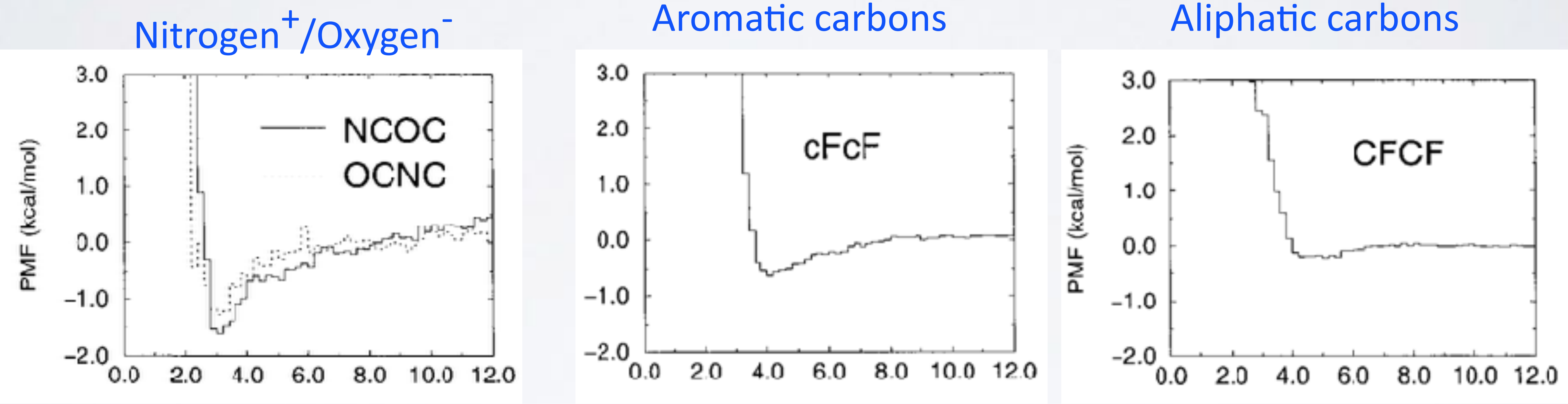
Find all protein-ligand structures in the PDB with a ligand carboxylate O

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



Atom-atom distance (Angstroms)

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

KNOWLEDGE-BASED POTENTIALS

Weaknesses

Accuracy limited by availability of data

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)

NEXT UP:

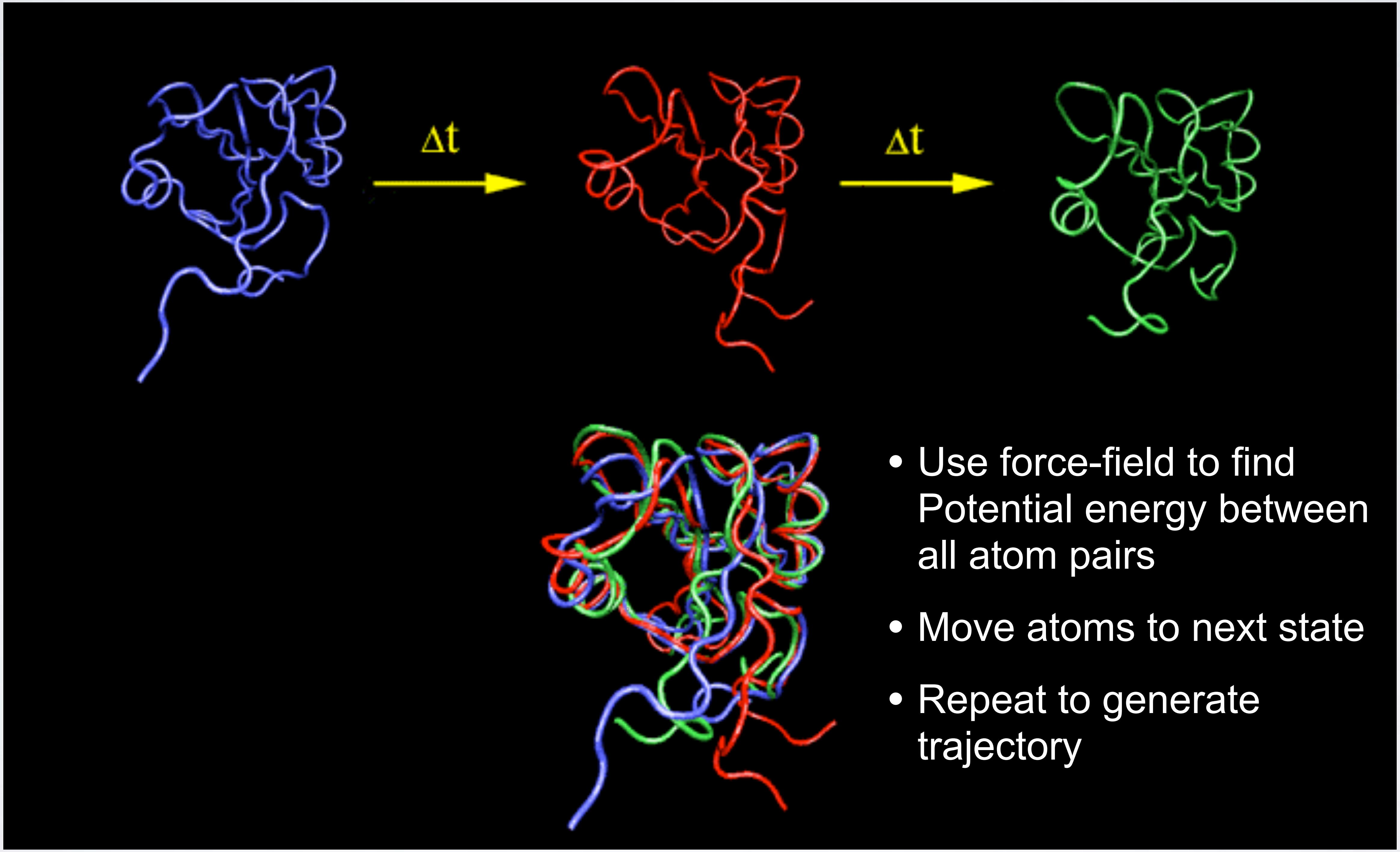
- **Overview of structural bioinformatics**
  - Major motivations, goals and challenges
- **Fundamentals of protein structure**
  - Composition, form, forces and dynamics
- **Representing and interpreting protein structure**
  - Modeling energy as a function of structure
- **Example application areas**
  - Predicting functional dynamics & drug discovery

PREDICTING FUNCTIONAL DYNAMICS

- **Proteins are intrinsically flexible molecules with internal motions that are often intimately coupled to their biochemical function**
  - E.g. ligand and substrate binding, conformational activation, allosteric regulation, etc.

- **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 and their properties

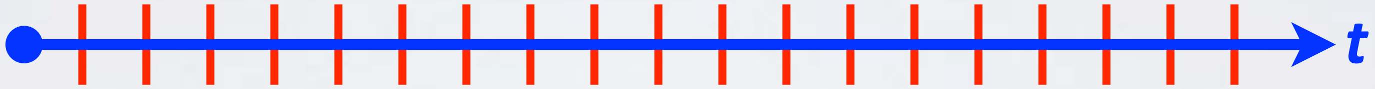
MOLECULAR DYNAMICS SIMULATION



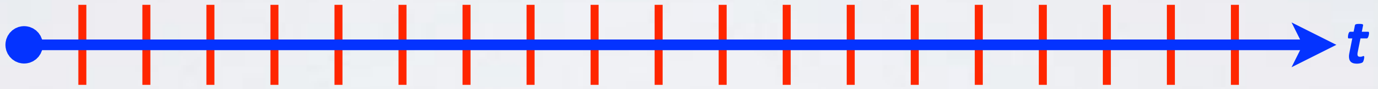
- Use force-field to find Potential energy between all atom pairs
- Move atoms to next state
- Repeat to generate trajectory

McCammon, Gelin & Karplus, *Nature* (1977)
[See: https://www.youtube.com/watch?v=ui1ZysMFcKk]

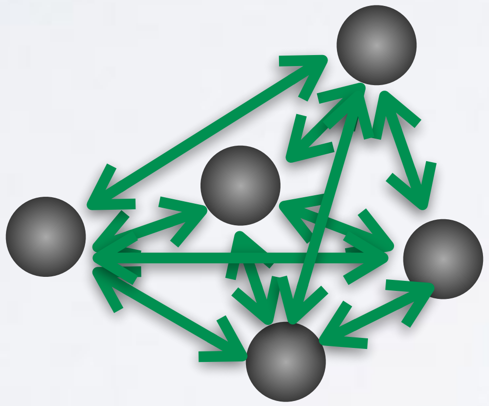
- Divide **time** into discrete (~1fs) **time steps** (**Δt**)
(for integrating equations of motion, see below)



- Divide **time** into discrete (~1fs) **time steps** (**Δt**)
  (for integrating equations of motion, see below)



- At each time step calculate pair-wise atomic **forces** (***F(t)***)
  (by evaluating **force-field** gradient)



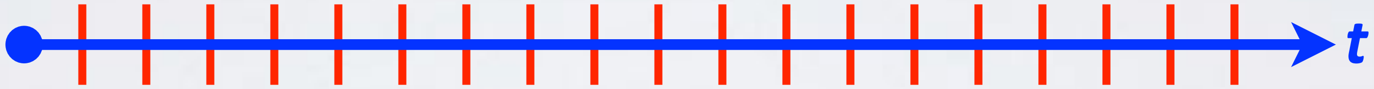
Nucleic motion described classically

$$m_i \frac{d^2}{dt^2}\vec{R}_i = -\vec{\nabla}_i E(\vec{R})$$

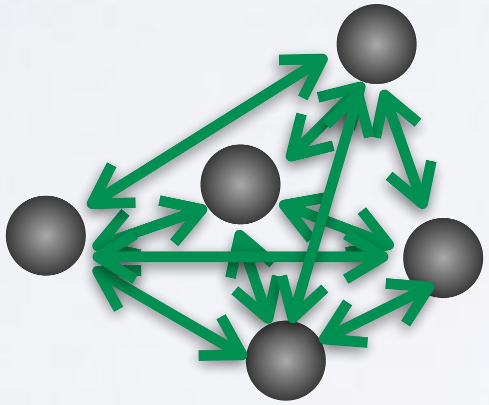
Empirical force field

$$E(\vec{R}) = \sum_{\text{bonded}} E_i(\vec{R}) + \sum_{\text{non-bonded}} E_i(\vec{R})$$

- Divide time into discrete (~1fs) time steps (Δt)
(for integrating equations of motion, see below)



- At each time step calculate pair-wise atomic forces (*F(t)*)
(by evaluating **force-field** gradient)



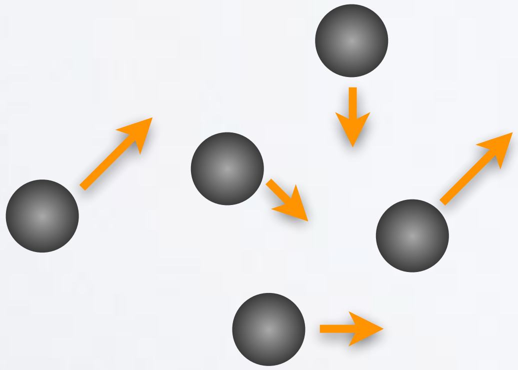
Nucleic motion described classically

$$m_i \frac{d^2}{dt^2}\vec{R}_i = -\vec{\nabla}_i E(\vec{R})$$

Empirical force field

$$E(\vec{R}) = \sum_{\mathrm{bonded}} E_i(\vec{R}) + \sum_{\mathrm{non-bonded}} E_i(\vec{R})$$

- Use the forces to calculate **velocities** and move atoms to new **positions**
(by integrating numerically via the "leapfrog" scheme)

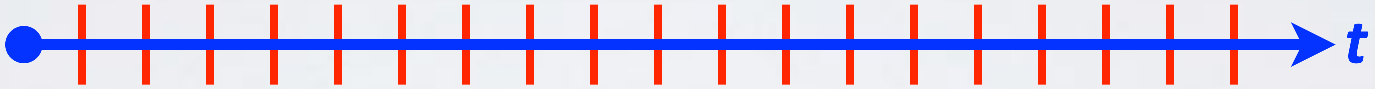


$$\boxed{\boldsymbol{v}(t + \frac{\Delta t}{2})} = \boldsymbol{v}(t - \frac{\Delta t}{2}) + \frac{\boldsymbol{F}(t)}{m}\Delta t$$

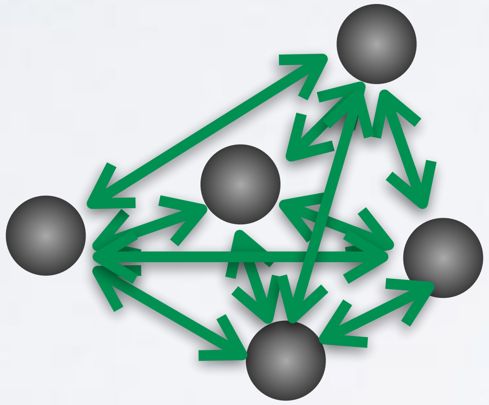
$$\boldsymbol{r}(t + \Delta t) = \boldsymbol{r}(t) + \boxed{\boldsymbol{v}(t + \frac{\Delta t}{2})}\Delta t$$

BASIC ANATOMY OF A MD SIMULATION

- Divide **time** into discrete (~1fs) **time steps** (**Δt**)
  (for integrating equations of motion, see below)



- At each time step calculate pair-wise atomic **forces** (***F(t)***)
  (by evaluating **force-field** gradient)



Nucleic motion described classically

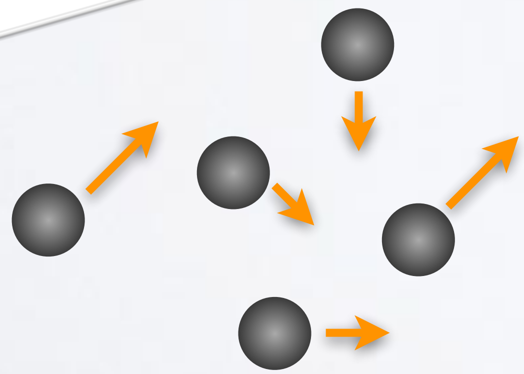
$$m_i \frac{d^2}{dt^2}\vec{R}_i = -\vec{\nabla}_i E(\vec{R})$$

Empirical f

$$E(\vec{R}) = \quad \sum_{\text{non-bonded}} E_i(\vec{R})$$

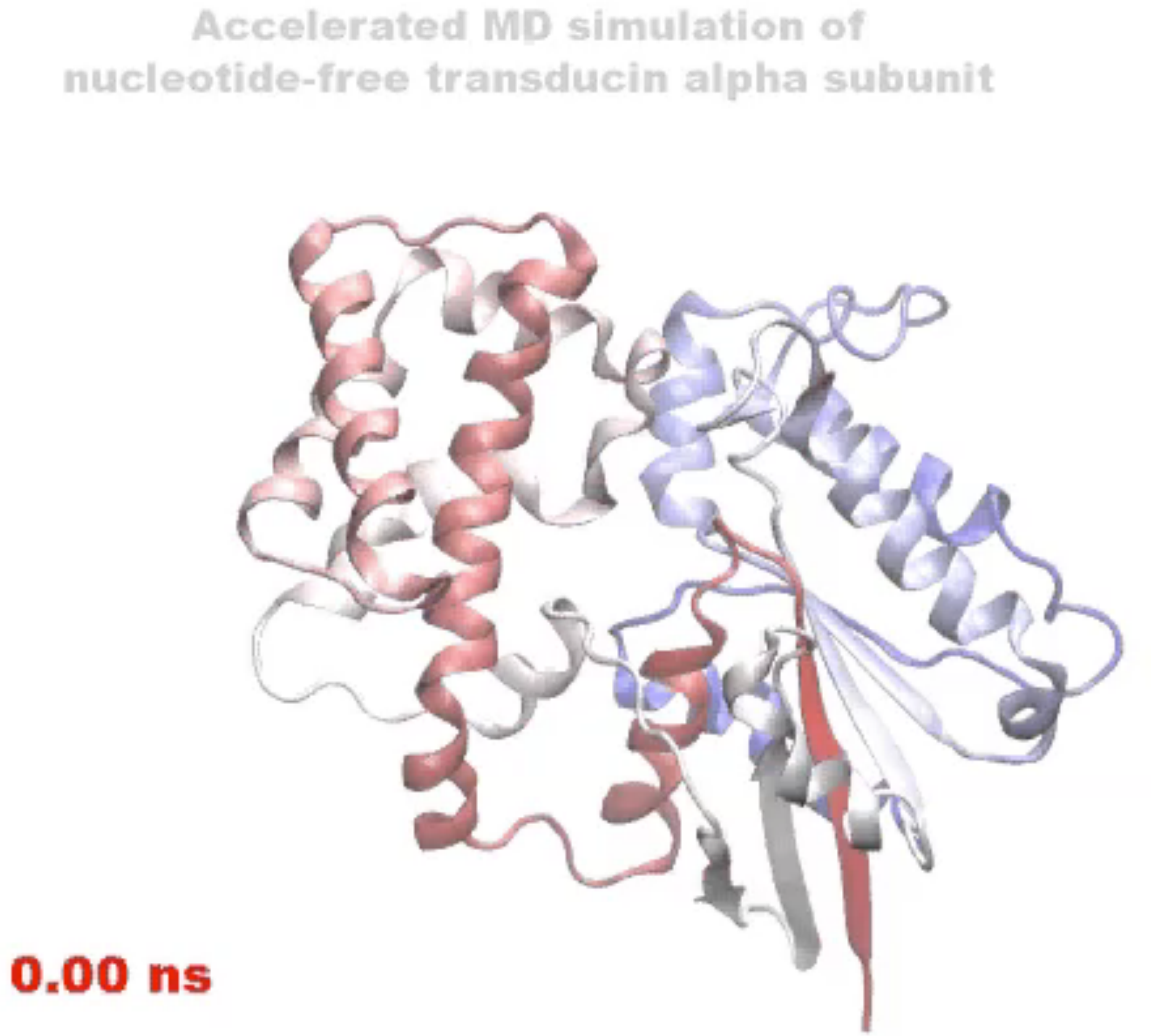
- Use th f ulate **velocities** and move atoms to new **positions**
  ing numerically via the “leapfrog” scheme)

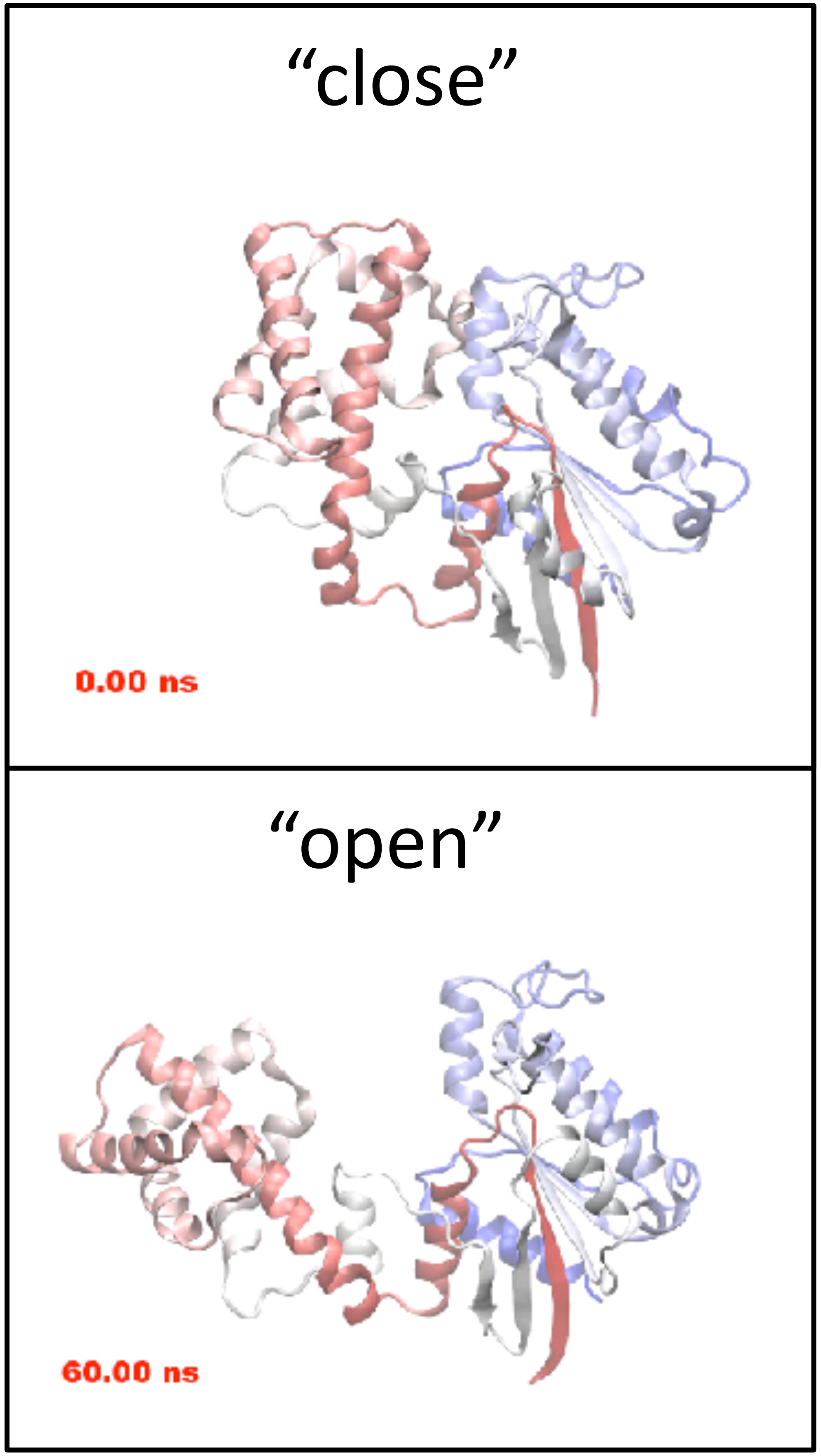
- **REPEAT, (iterate many, many times... 1ms = 10^{12} time steps)**



$$\begin{aligned} \boldsymbol{v}(t+\frac{\Delta t}{2}) &= \boldsymbol{v}(t-\frac{\Delta t}{2}) + \frac{\boldsymbol{F}(t)}{m}\Delta t \\ \boldsymbol{r}(t+\Delta t) &= \boldsymbol{r}(t) + \boldsymbol{v}(t+\frac{\Delta t}{2})\Delta t \end{aligned}$$

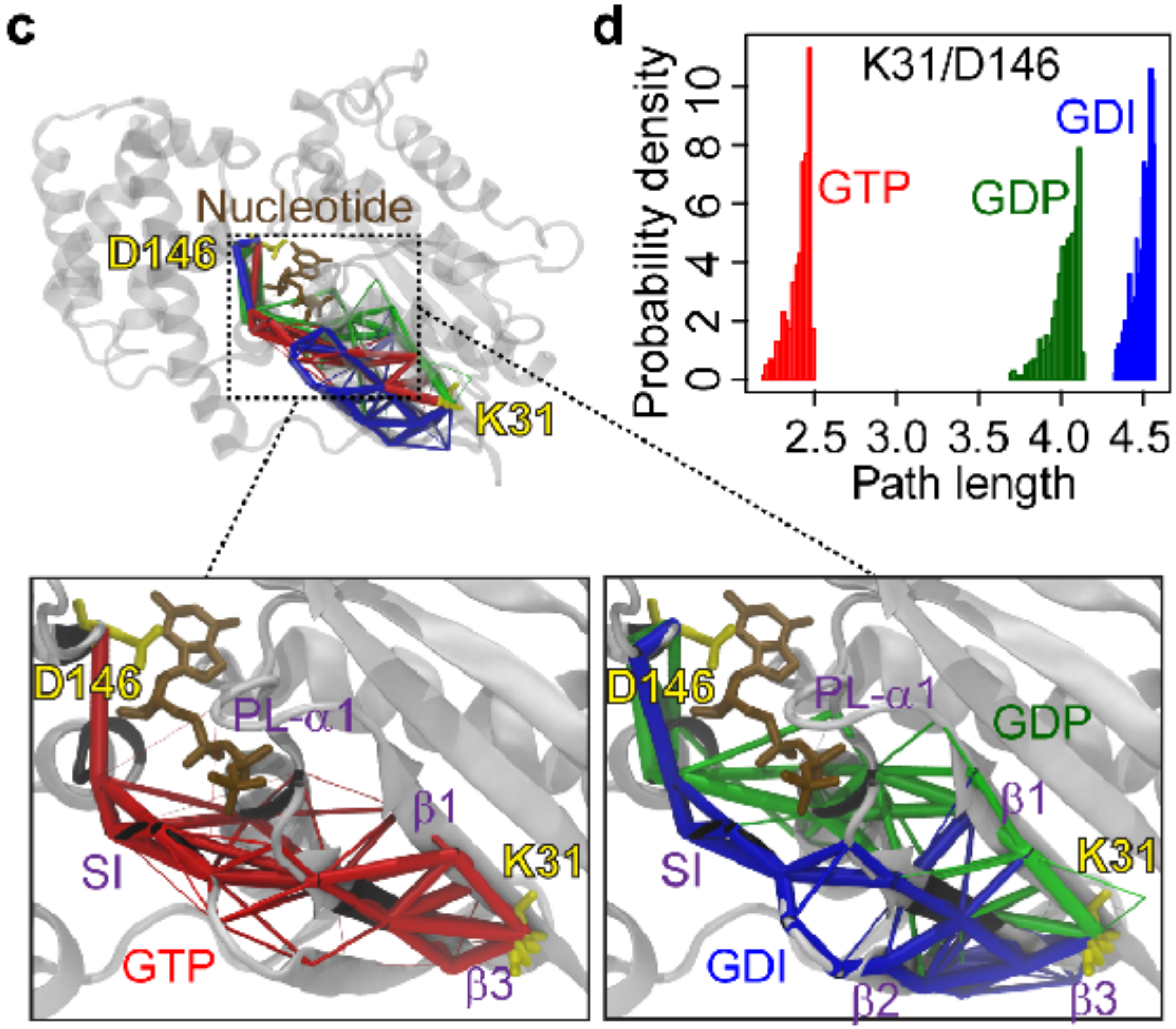
MD Prediction of Functional Motions


Accelerated MD simulation of
nucleotide-free transducin alpha subunit
0.00 ns



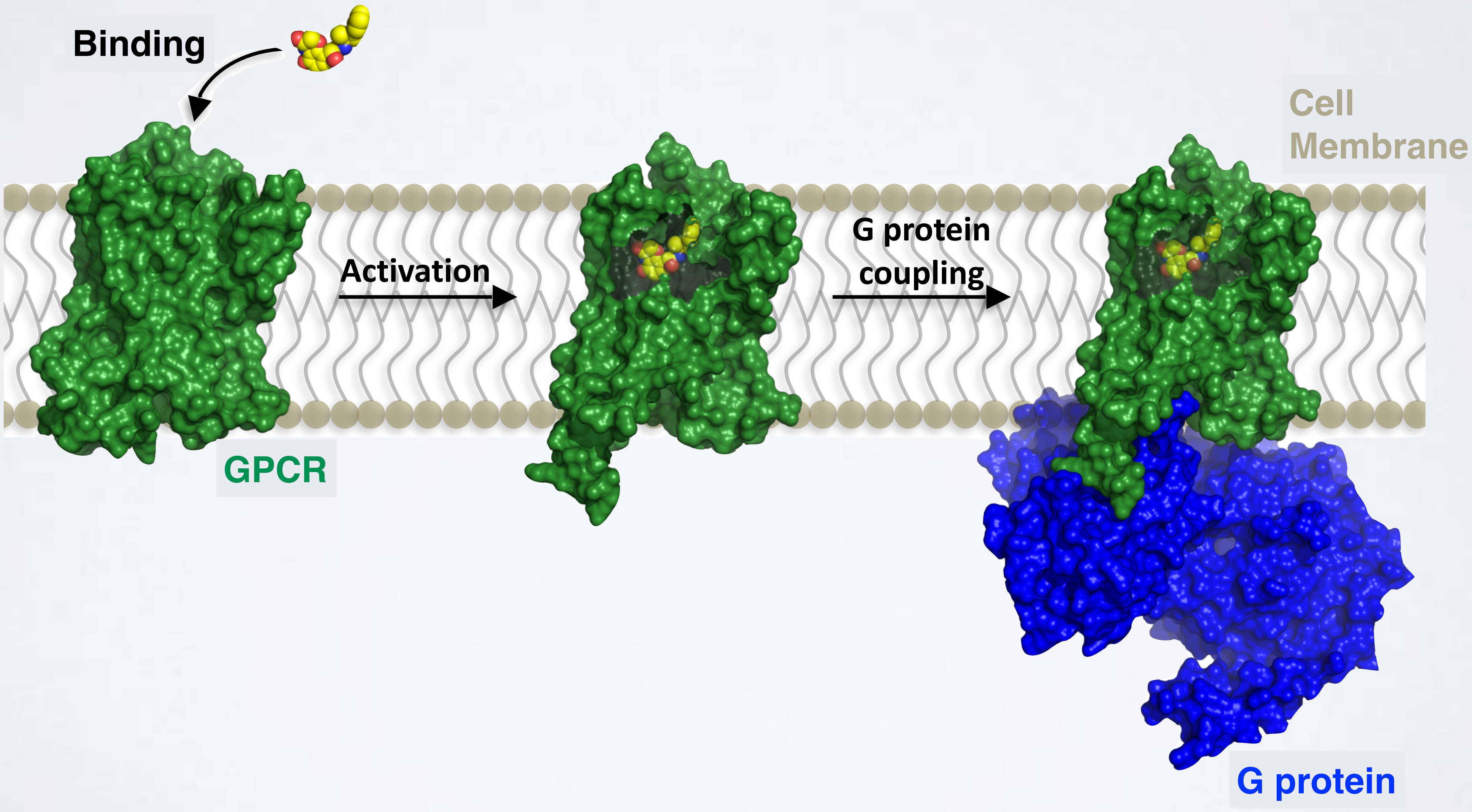
Yao and Grant, Biophys J. (2013)

Simulations Identify Key Residues Mediating Dynamic Activation

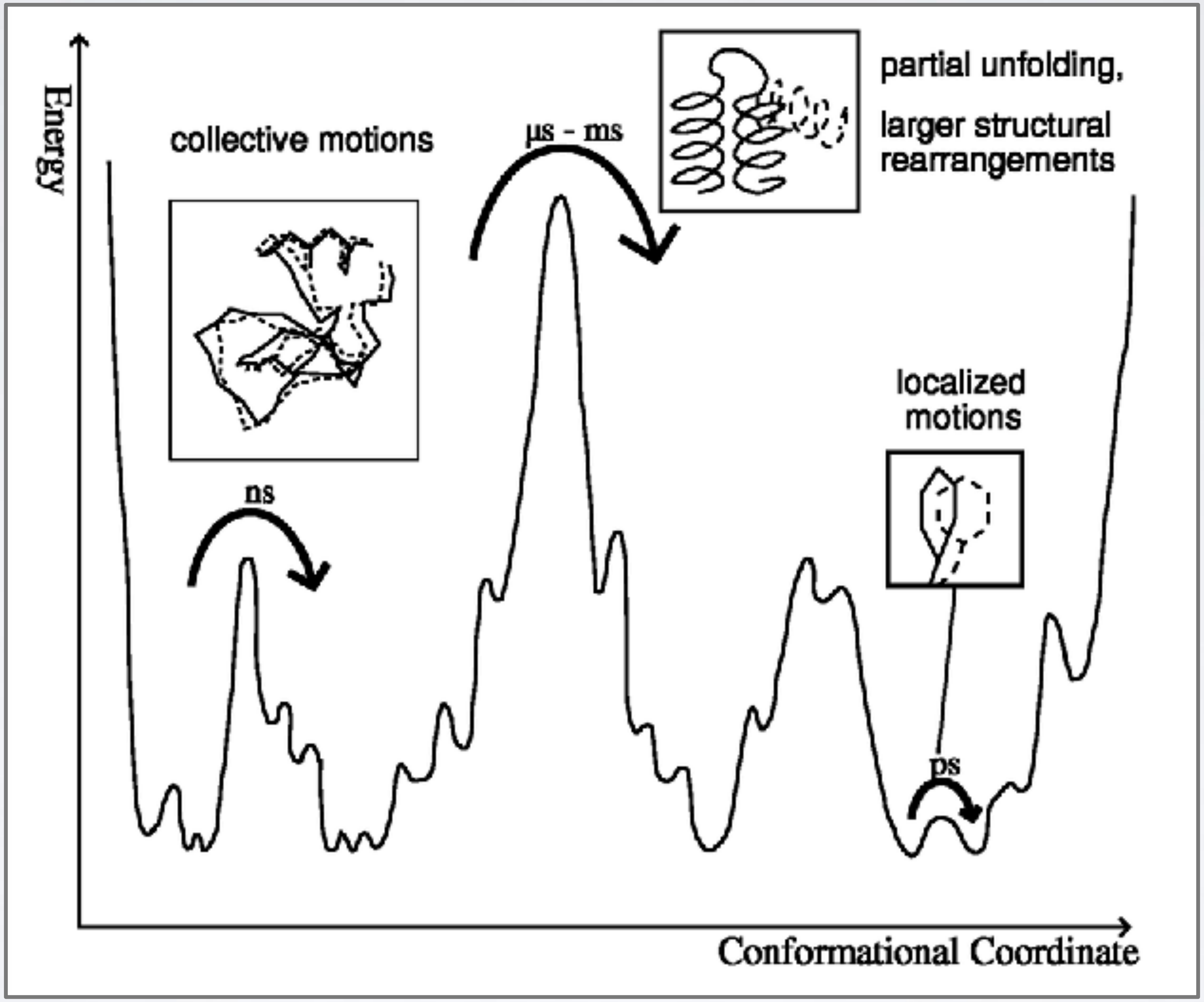


Yao … Grant, Journal of Biological Chemistry (2016)

EXAMPLE APPLICATION OF MOLECULAR SIMULATIONS TO GPCRS



PROTEINS JUMP BETWEEN MANY, HIERARCHICALLY ORDERED "CONFORMATIONAL SUBSTATES"



H. Frauenfelder et al., *Science* **229** (1985) 337

Improve this slide

MOLECULAR DYNAMICS IS VERY

Example: F_1-ATPase in water (183,674 atoms) for 1 nanosecond:

=> 10^6 integration steps

=> $8.4 * 10^{11}$ floating point operations/step

[n(n-1)/2 interactions]

Total: $8.4 * 10^{17}$ flop

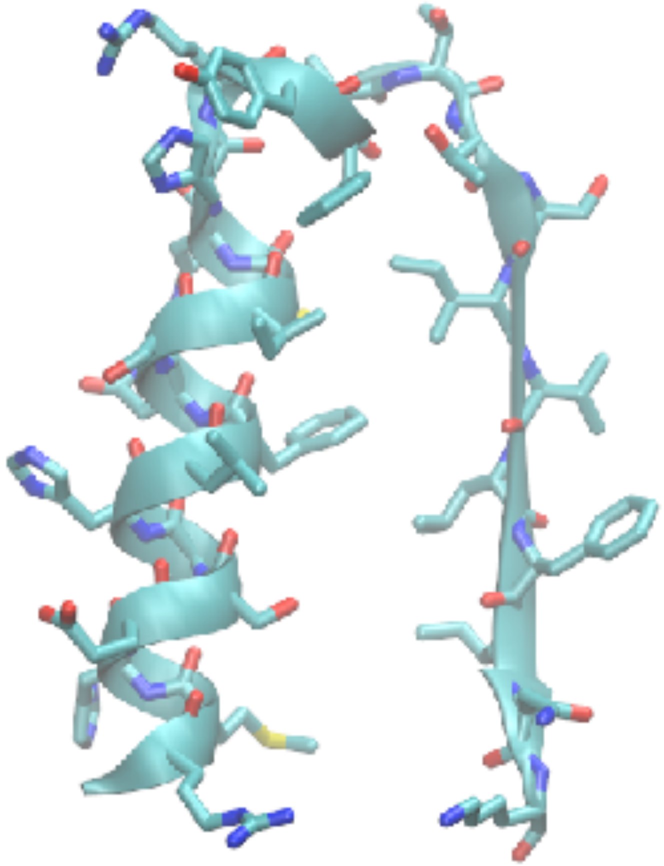
(on a 100 Gflop/s cpu: **ca 25 years!**)

… but performance has been improved by use of:

multiple time stepping	ca. 2.5 years
fast multipole methods	ca. 1 year
parallel computers	ca. 5 days
modern GPUs	**ca. 1 day**
(Anton supercomputer	**ca. minutes)**

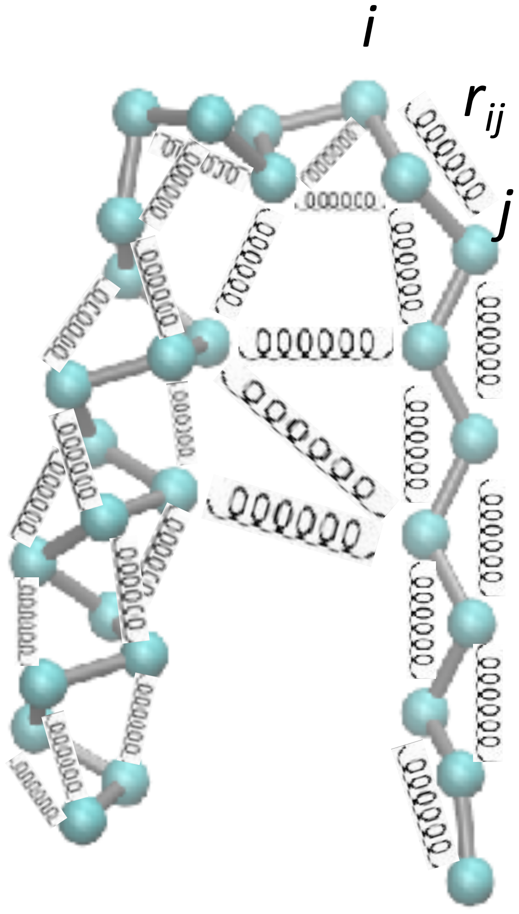
COARSE GRAINING: **NORMAL MODE ANALYSIS** (NMA)

- MD is still time-consuming for large systems
- Elastic network model NMA (ENM-NMA) is an example of a lower resolution approach that finishes in seconds even for large systems.



Atomistic

C. G.



Coarse Grained

- 1 bead / 1 amino acid
- Connected by springs

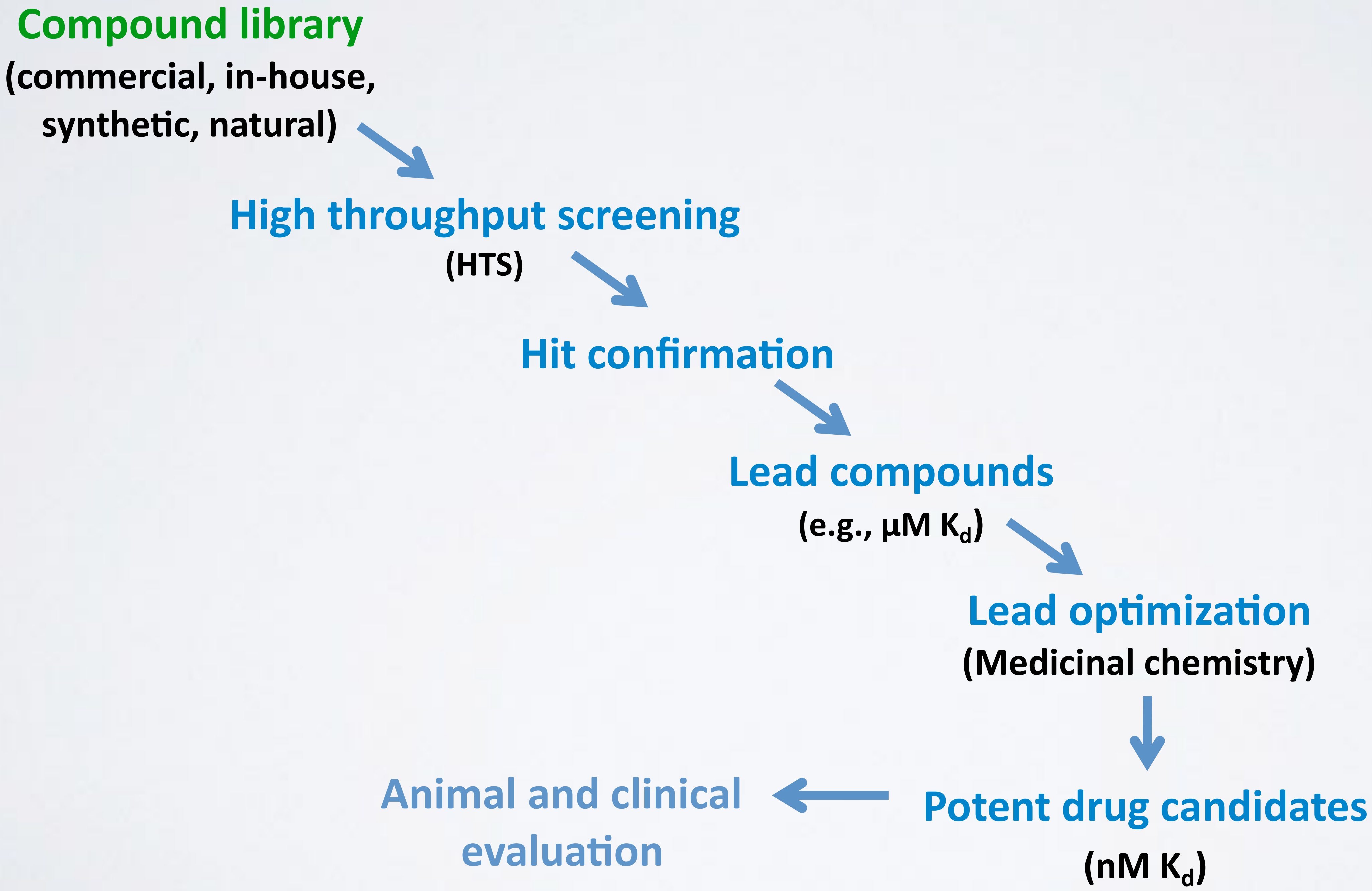
NMA models the protein as a network of elastic strings

Proteinase K

NEXT UP:

- **Overview of structural bioinformatics**
  - Major motivations, goals and challenges
- **Fundamentals of protein structure**
  - Composition, form, forces and dynamics
- **Representing and interpreting protein structure**
  - Modeling energy as a function of structure
- **Example application areas**
  - Predicting functional dynamics & **drug discovery**

THE TRADITIONAL EMPIRICAL PATH TO DRUG DISCOVERY



Compound library

(commercial, in-house, synthetic, natural)

High throughput screening

(HTS)

Hit confirmation

Lead compounds

(e.g., µM K_d)

Lead optimization

(Medicinal chemistry)

Potent drug candidates

(nM K_d)

Animal and clinical evaluation

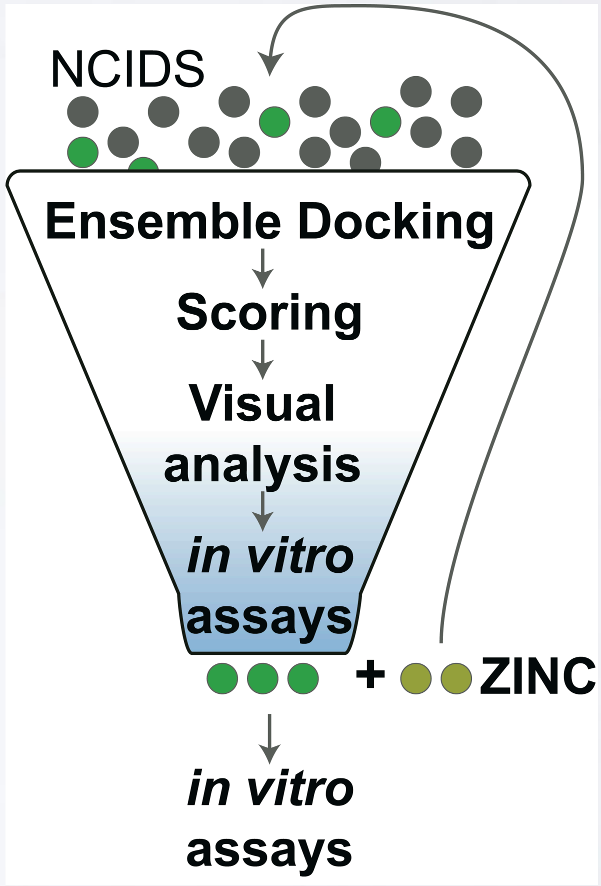
COMPUTER-AIDED LIGAND DESIGN

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**

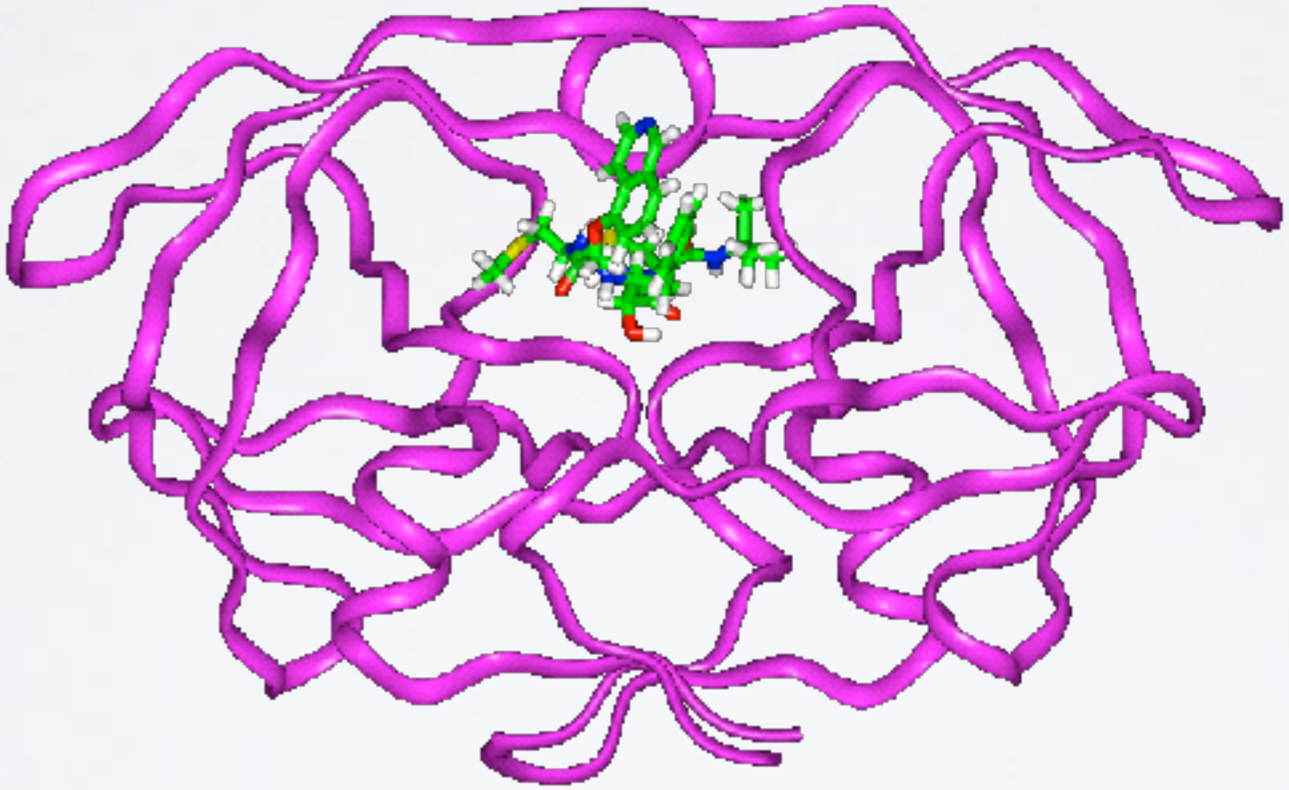
Two main approaches:

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

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

SCENARIO I: RECEPTOR-BASED DRUG DISCOVERY

Structure of Targeted Protein Known: Structure-Based Drug Discovery



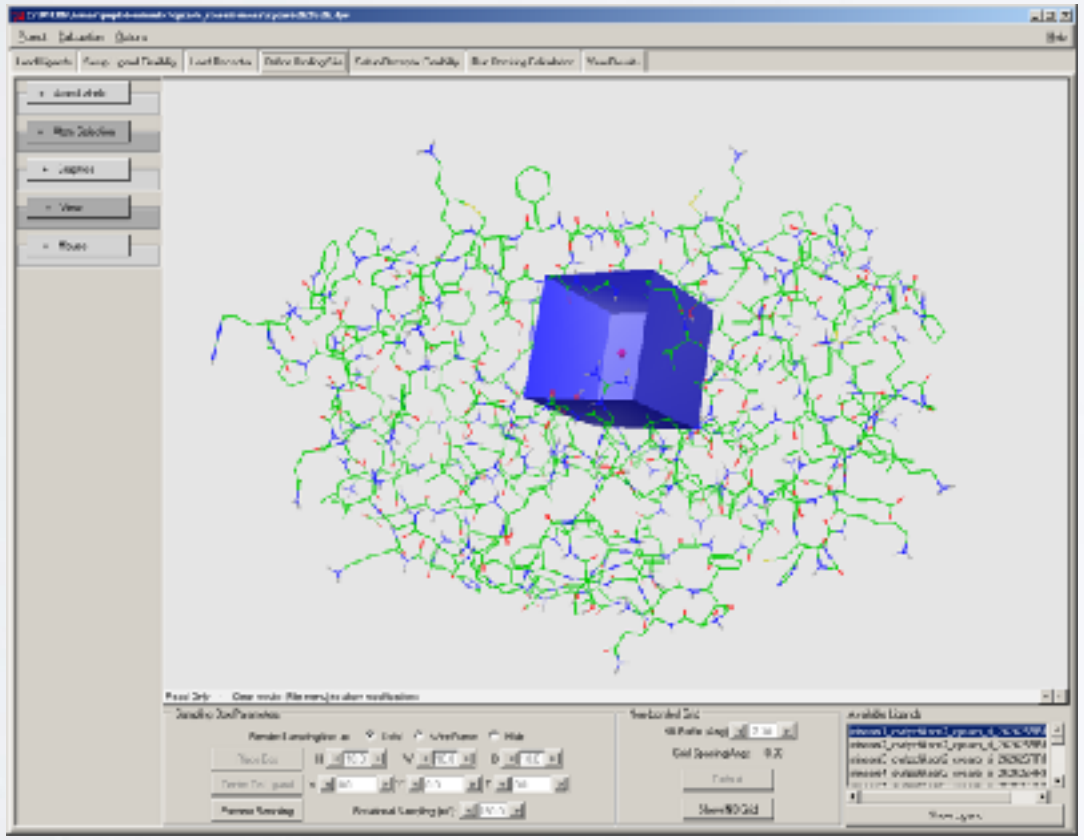
HIV Protease/KNI-272 complex

PROTEIN-LIGAND DOCKING

Structure-Based Ligand Design

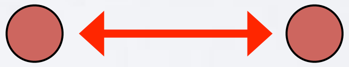
Docking software

Search for structure of lowest energy

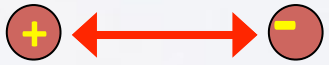


Potential function

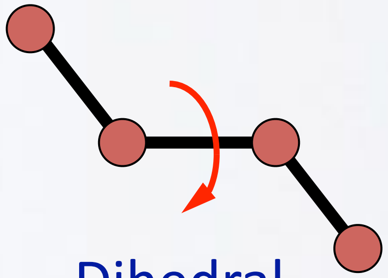
Energy as function of structure



VDW

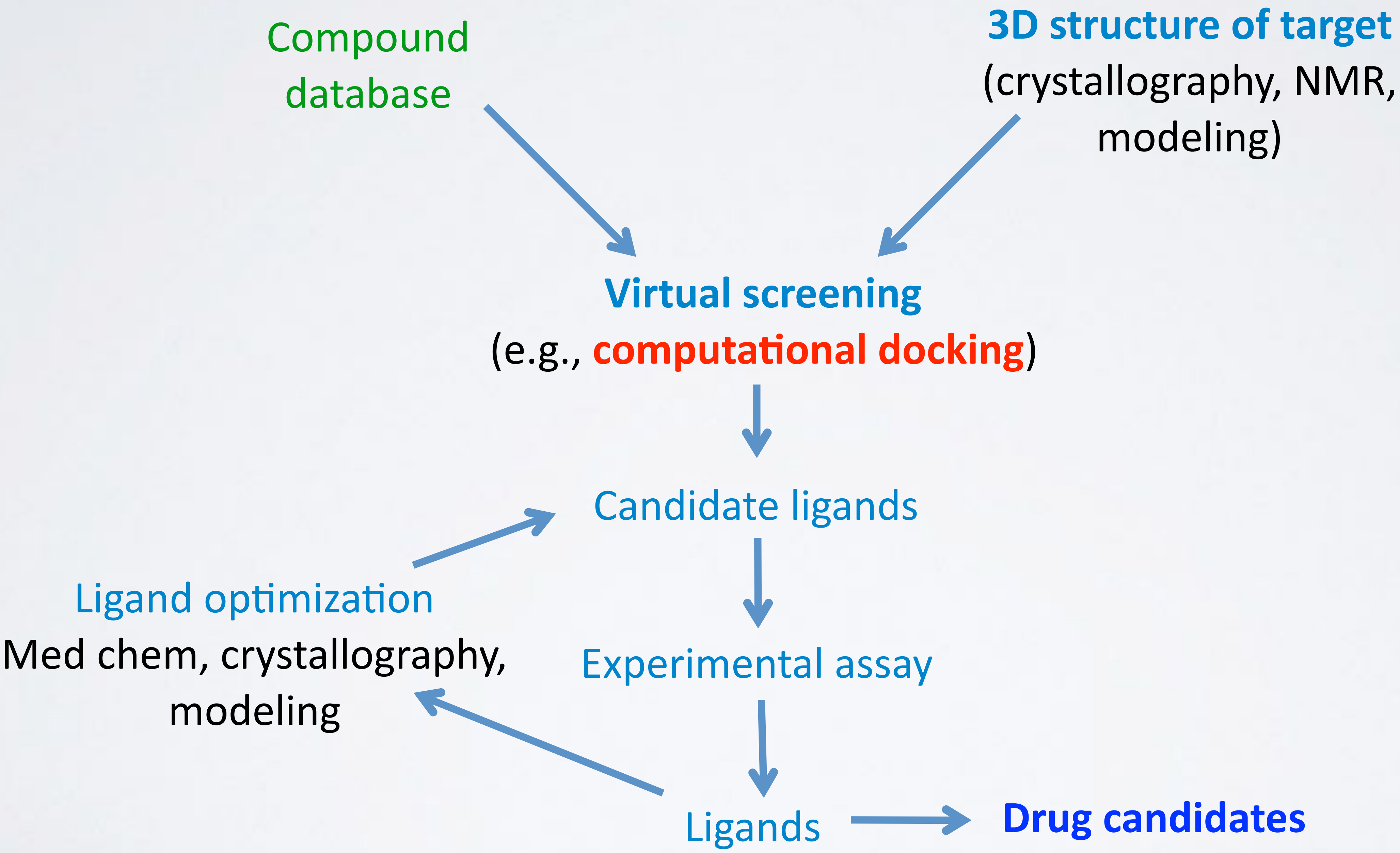


Screened Coulombic



Dihedral

STRUCTURE-BASED VIRTUAL SCREENING



Compound database

3D structure of target
(crystallography, NMR, modeling)

Virtual screening
(e.g., **computational docking**)

Candidate ligands

Experimental assay

Ligands

Drug candidates

Ligand optimization
Med chem, crystallography, modeling

COMPOUND LIBRARIES

Maybridge HitFinder™

This pre-selected diverse screening library makes identifying potential drug leads easy, convenient and cost effective.

[illegible]

Commercial
(in-house pharma)

NIH MOLECULAR LIBRARIES
SMALL MOLECULE REPOSITORY

BioFocus
A Galapagos Company

A Roadmap Initiative

- Home
- MLSMR Project
  - Compound Identification
  - Quality Control
  - Sample Storage
  - Sample Arrays
  - Information
- MLPCN Centers
- MLSMR Contacts
- Submit Compounds

Registered Users Login

Copyright © 2007 Galapagos NV

Welcome

NIH Molecular Libraries Small Molecule Repository collects samples for high throughput biological screening and distributes them to the NIH Molecular Libraries Probe Production Centers Network. Learn more.

MLSMR is a key component of the Molecular Libraries Initiative, an NIH Roadmap project supporting New Pathways to Discovery in the 21st century. The project is funded in whole with Federal funds from the National Institutes of Health, Department of Health and Human Services, under Contract No. HHS-N-278-2004-41001C.

In the news:

Behind the Scenes at the NIH Molecular Libraries Small Molecule Repository

The NIH Molecular Libraries Small Molecule Repository is now selling the NIH Clinical Collection

BioFocus, a Galapagos company operates MLSMR in South San Francisco.

Government (NIH)

University of Pittsburgh

PITT HOME | FIND PEOPLE | CONTACT US

Pittsburgh Molecular Libraries Screening Center

- HOME
- HISTORY
- PERSONNEL
- SCREENING TECHNOLOGY
- COMPOUND LIBRARIES
- INSTRUMENTS & PLATFORMS
- HTS GUIDELINES
- APPROVED PMLSC ASSAY PROTOCOLS
- PMLSC PROBE REPORTS
- CHEMISTRY
- DATA ANALYSIS/INFORMATICS
- EDUCATIONAL ACTIVITIES
- PUBLICATIONS
- LINKS
- CONTACTS

Keyword Search

Go

Welcome

The Pittsburgh Molecular Library Screening Center (PMLSC) comprises investigators at the University of Pittsburgh and Carnegie Mellon University. Its mission is to assist scientists and the National Institutes of Health to thoughtfully interrogate small molecule libraries using optical-based High Throughput and High Content assays.

Health Sciences @ Pitt | UPMC | HSLS | School of Medicine | Health Sciences Calendar | Our News & Events

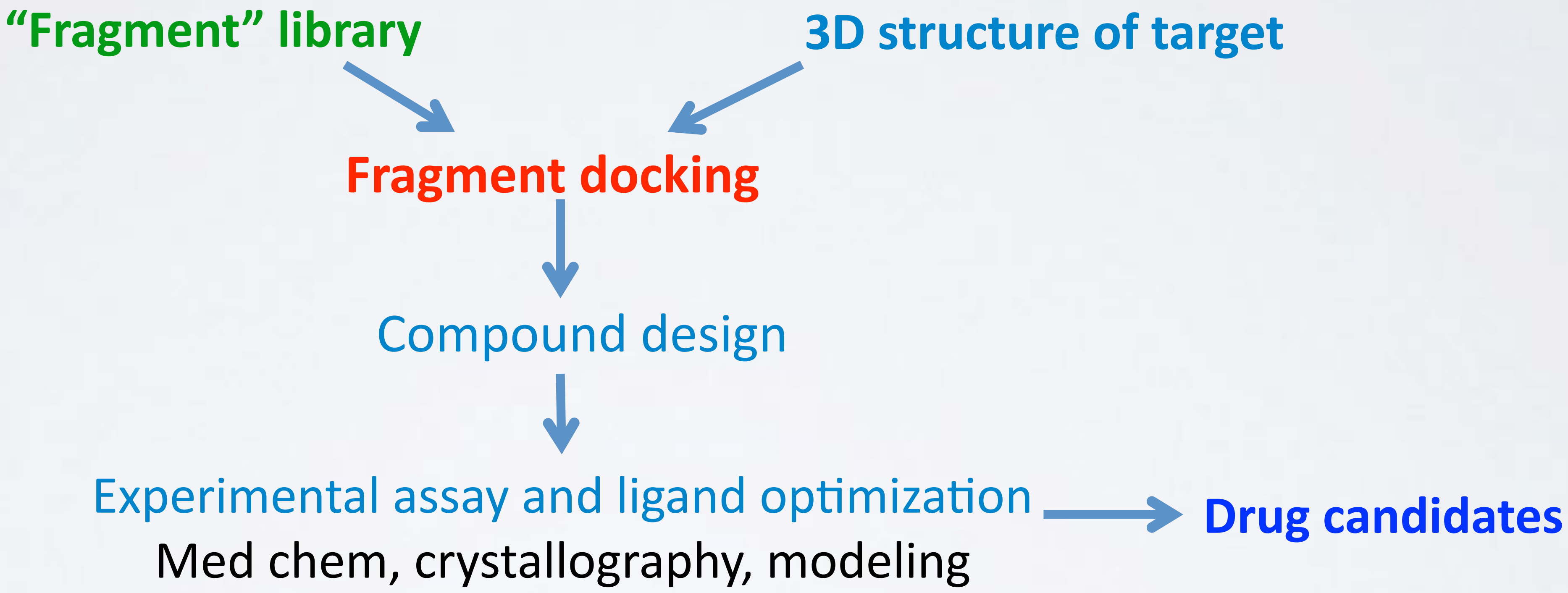
Top of Page | Home | Contact Us

Last Update [illegible]

© Office of the Senior Vice Chancellor for Health Sciences, University of Pittsburgh. All rights reserved.

Academia

FRAGMENTAL STRUCTURE-BASED SCREENING



"Fragment" library

3D structure of target

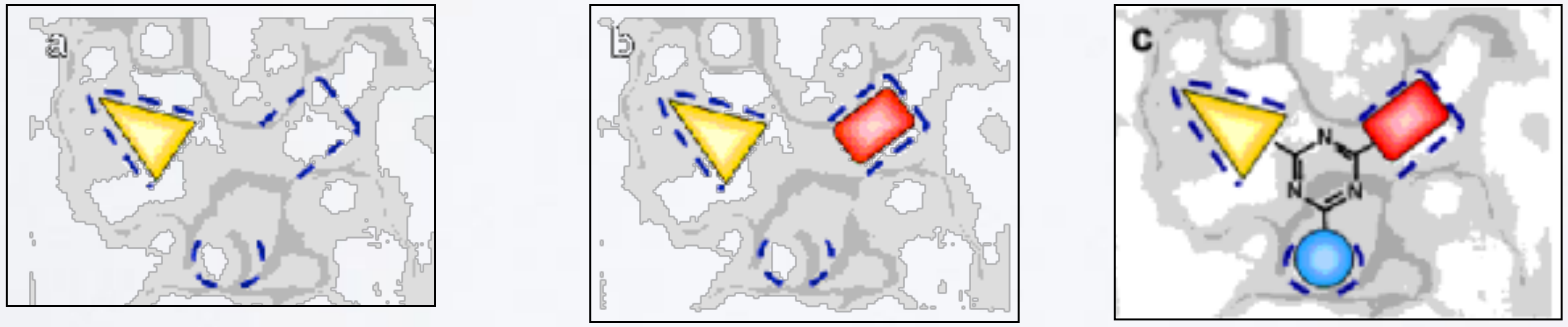
Fragment docking

Compound design

Experimental assay and ligand optimization

Med chem, crystallography, modeling

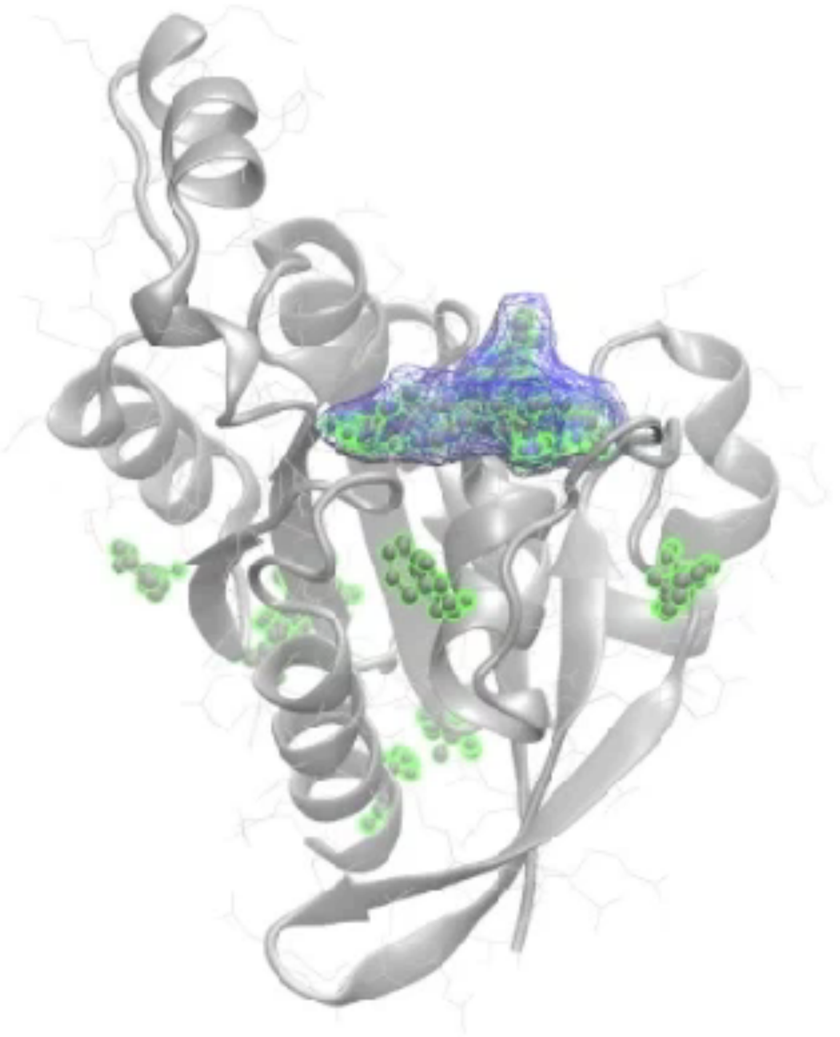
Drug candidates

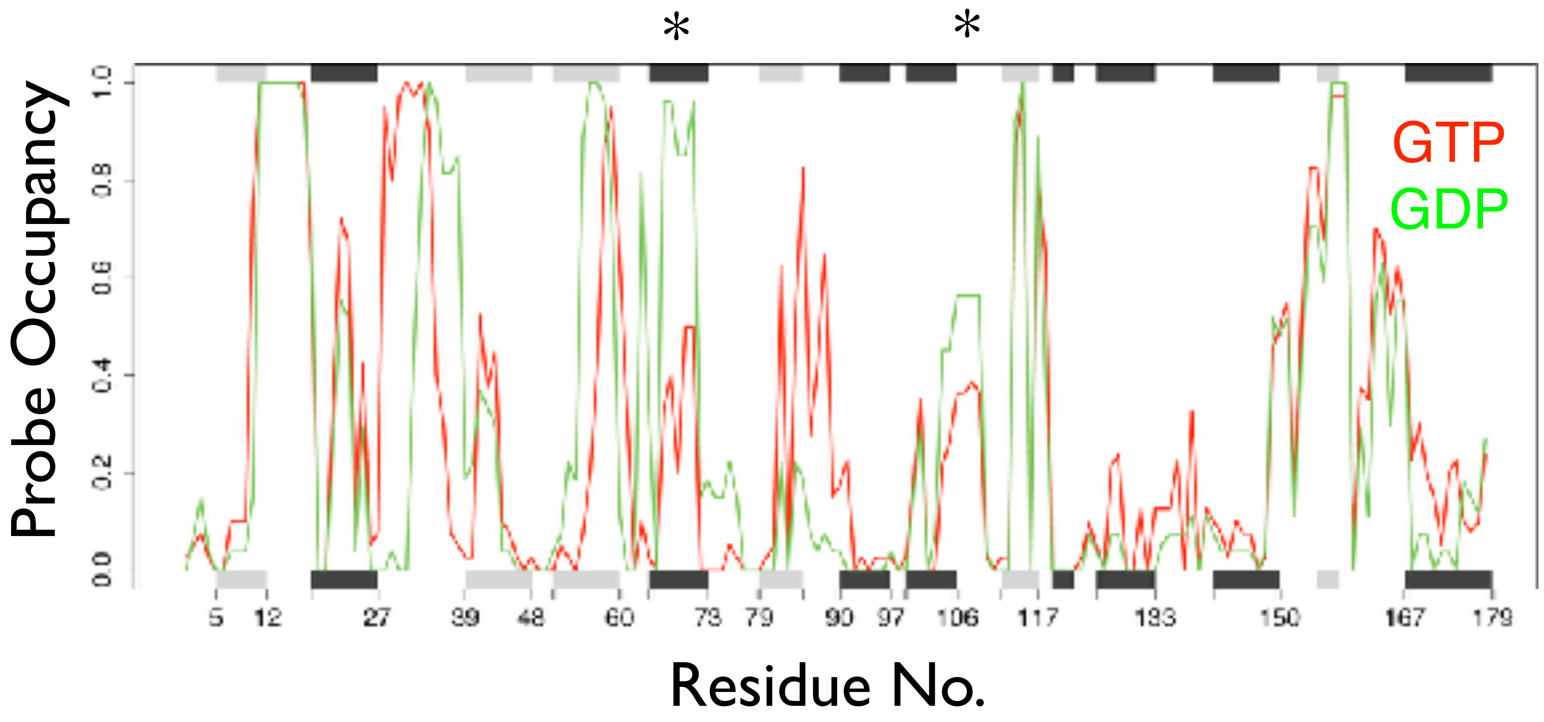


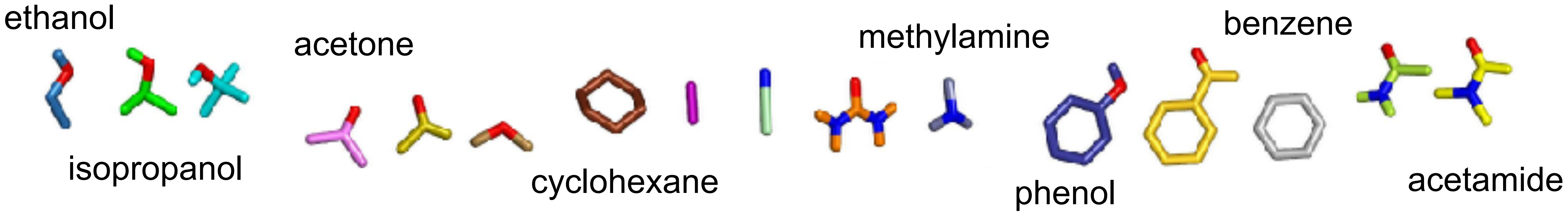
http://www.beilstein-institut.de/bozen2002/proceedings/Jhoti/jhoti.html

Multiple non active-site pockets identified

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



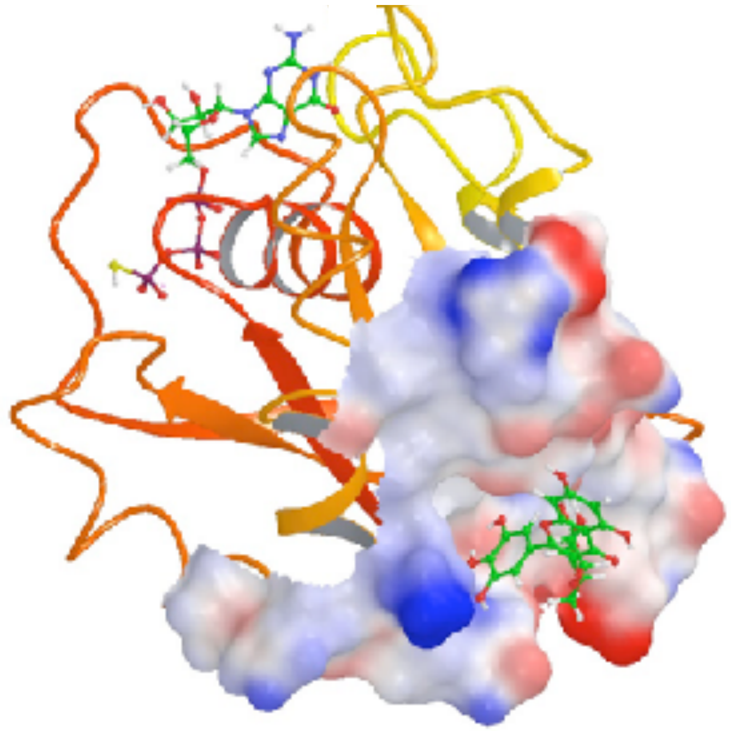


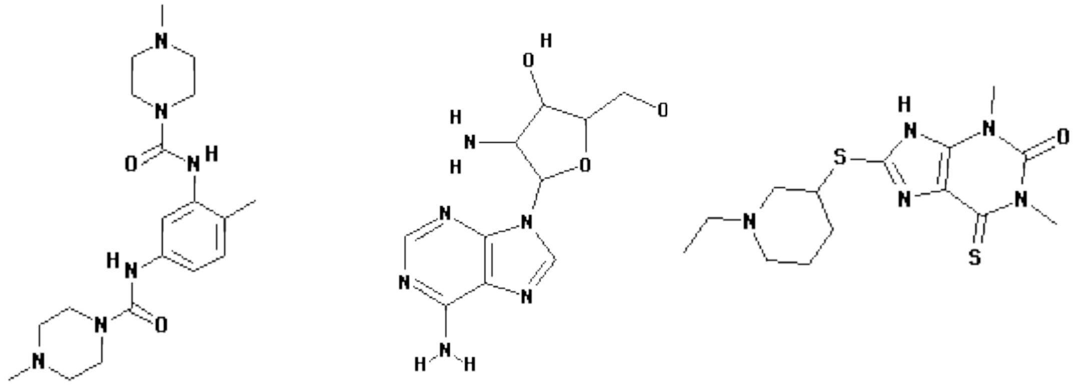


Ensemble docking & candidate inhibitor testing

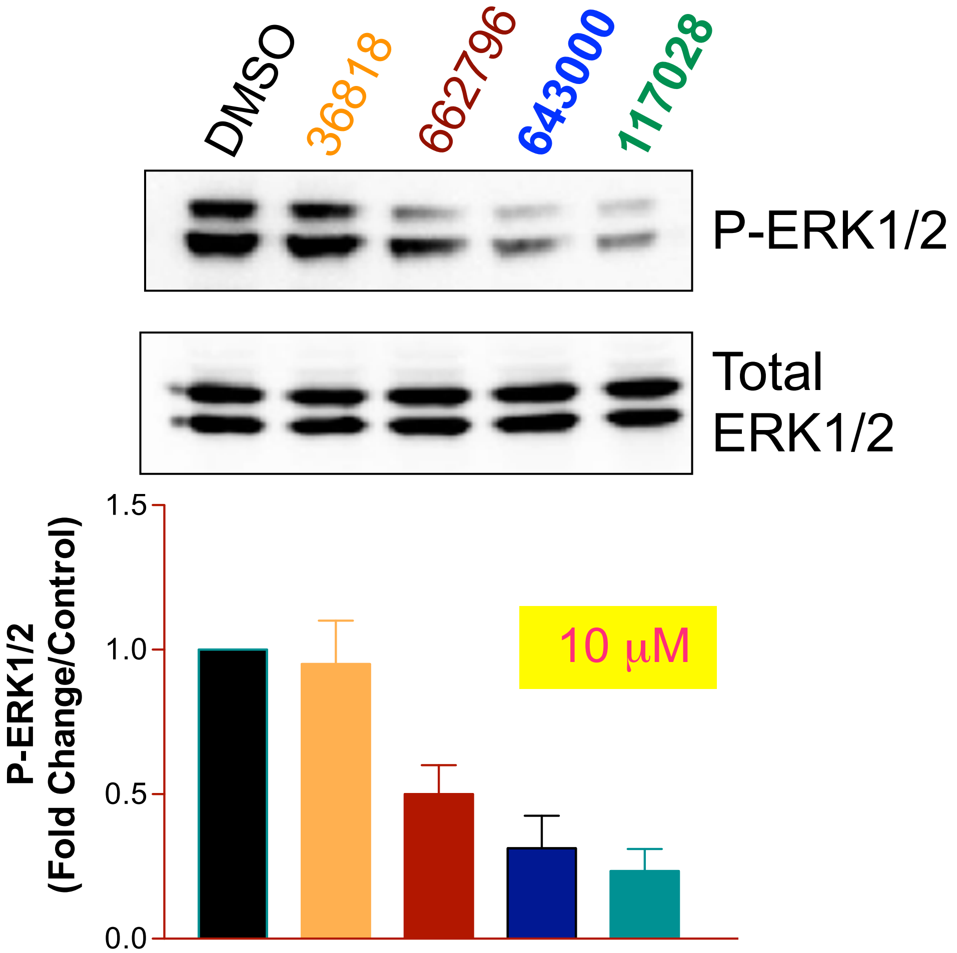
Top hits from ensemble docking against distal pockets were tested for inhibitory effects on basal ERK activity in glioblastoma cell lines.

Ensemble computational docking





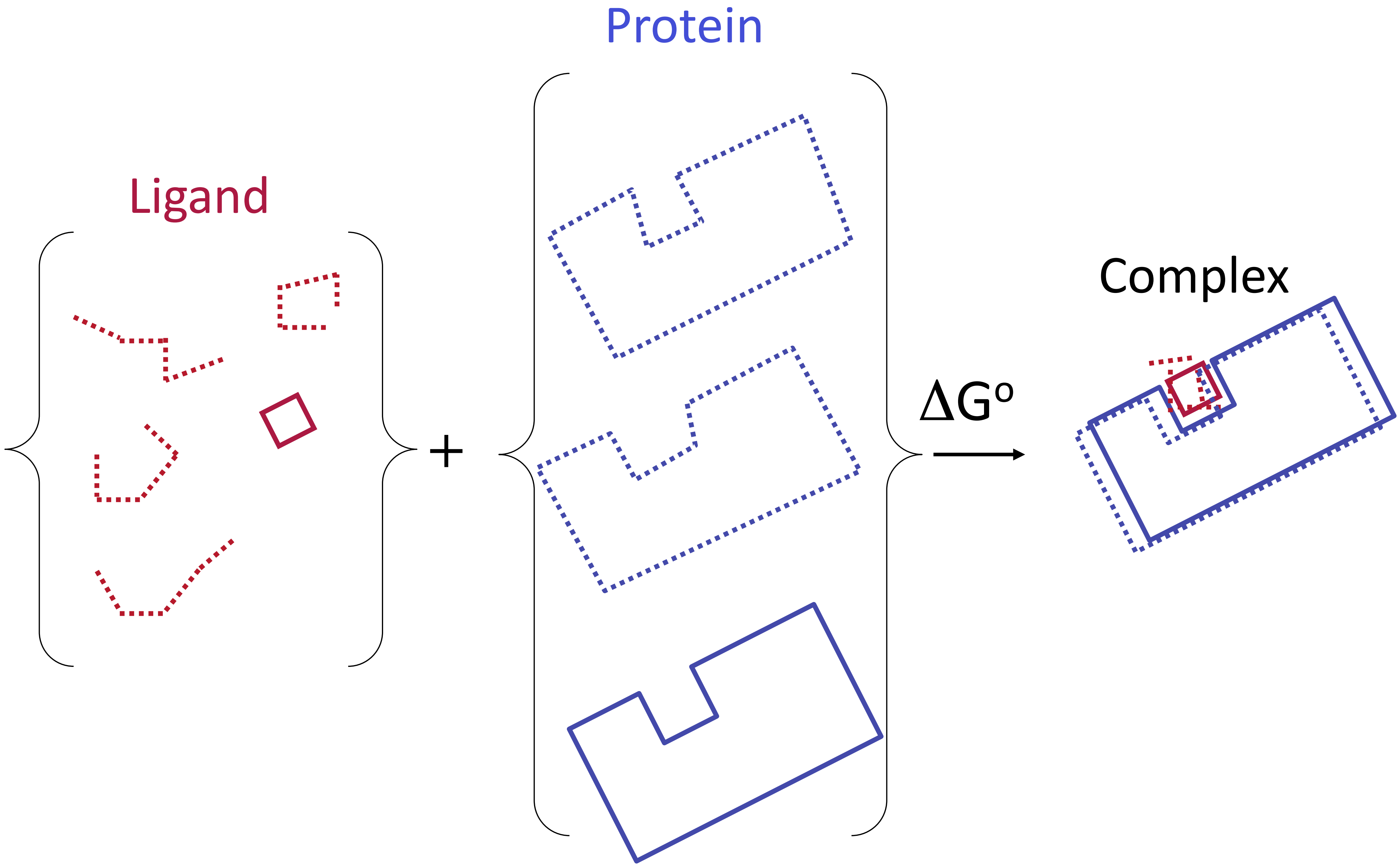
Compound effect on U251 cell line



Compound testing in cancer cell lines

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**

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

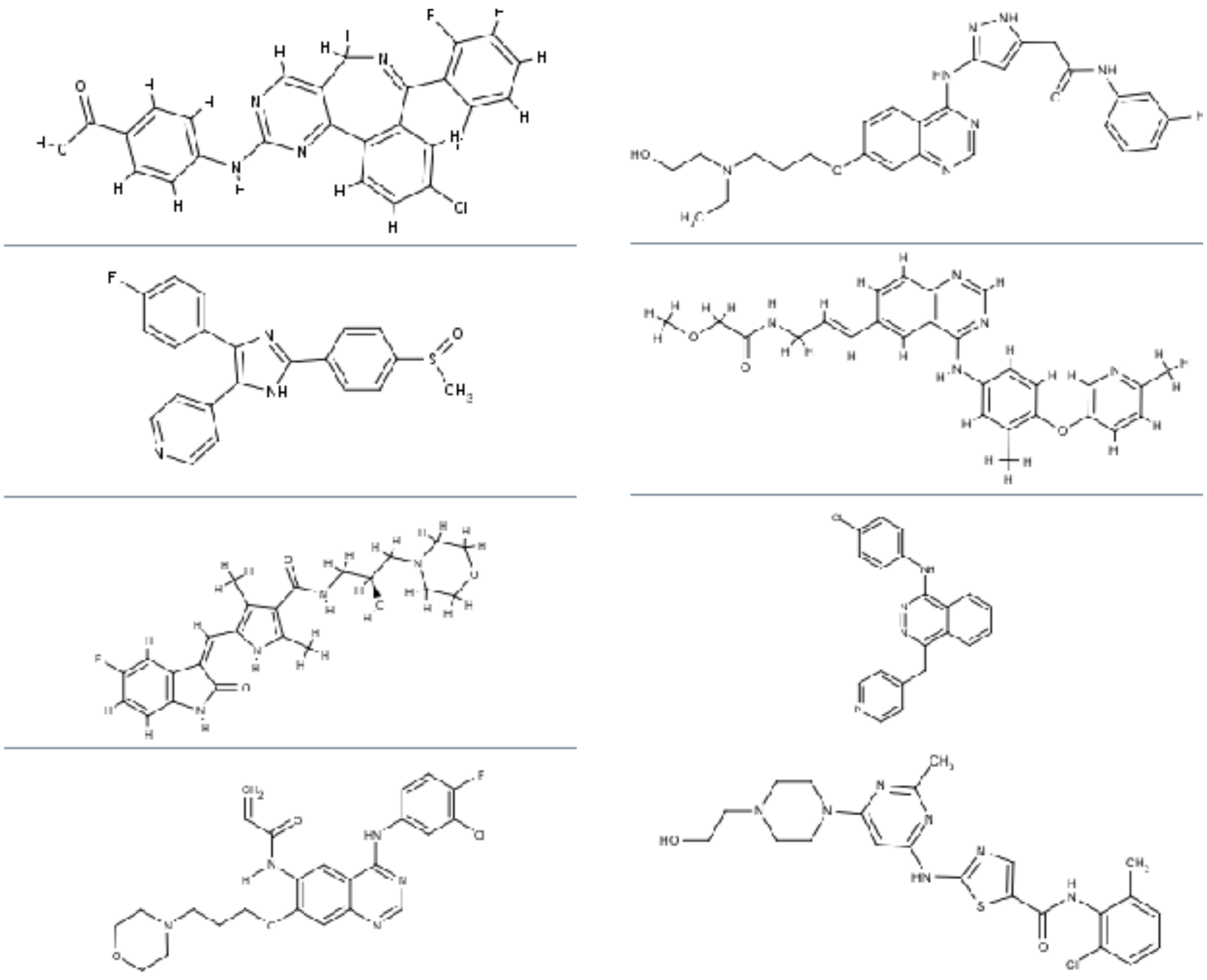
A company wants to work around another company’s chemical patents

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

Scenario 2

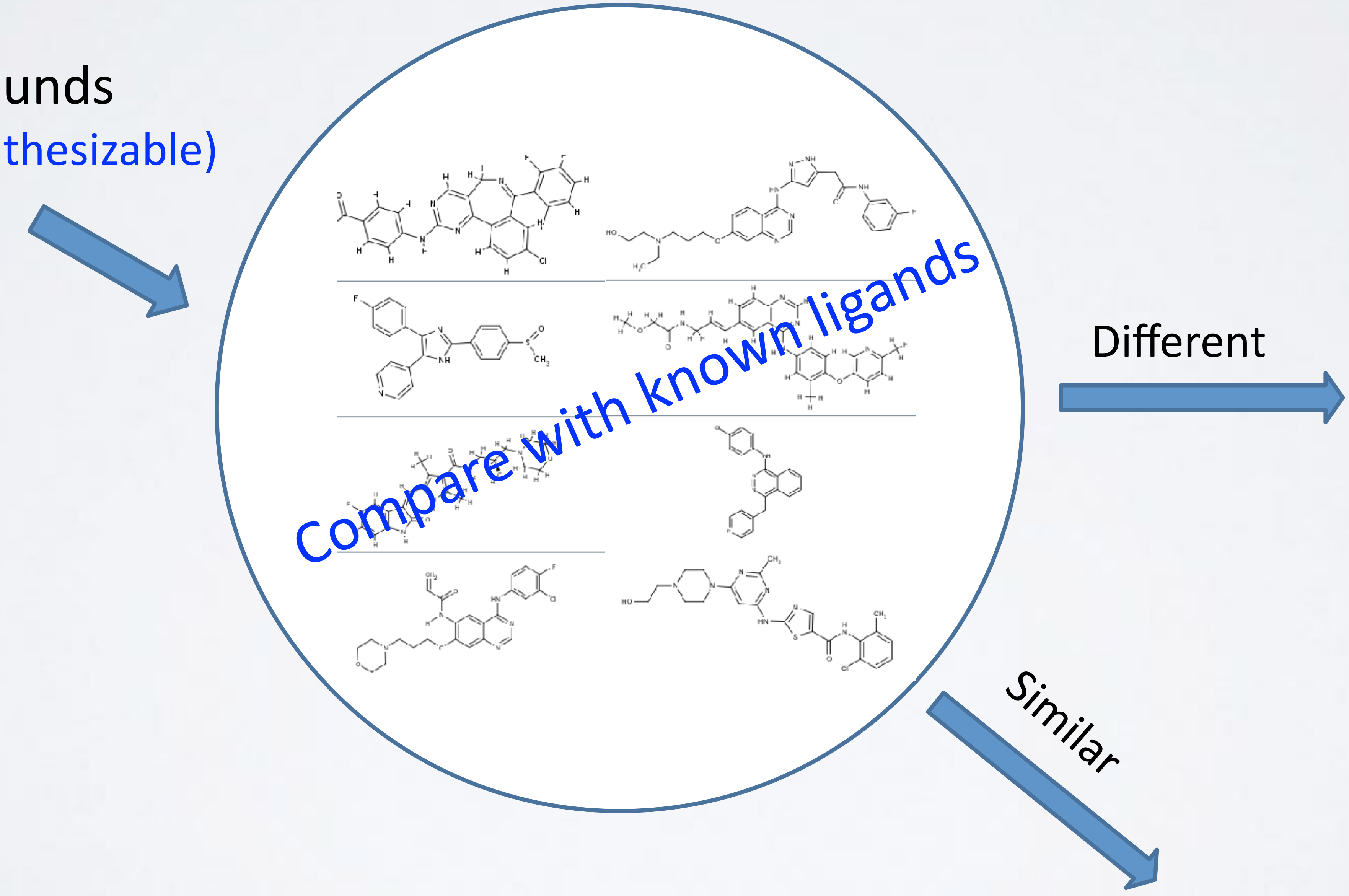
Structure of Targeted Protein Unknown: Ligand-Based Drug Discovery

e.g. MAP Kinase Inhibitors



Using knowledge of existing inhibitors to discover more

CHEMICAL SIMILARITY
LIGAND-BASED DRUG-DISCOVERY



Compounds
(available/synthesizable)

Compare with known ligands

Different

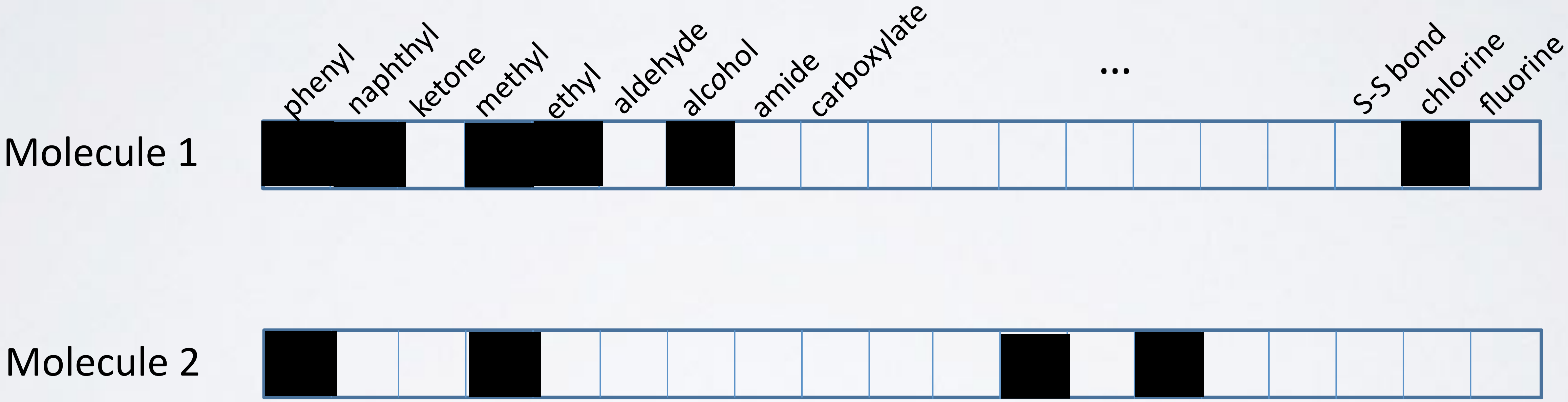
Don’t bother

Similar

Test experimentally

CHEMICAL FINGERPRINTS

BINARY STRUCTURE KEYS

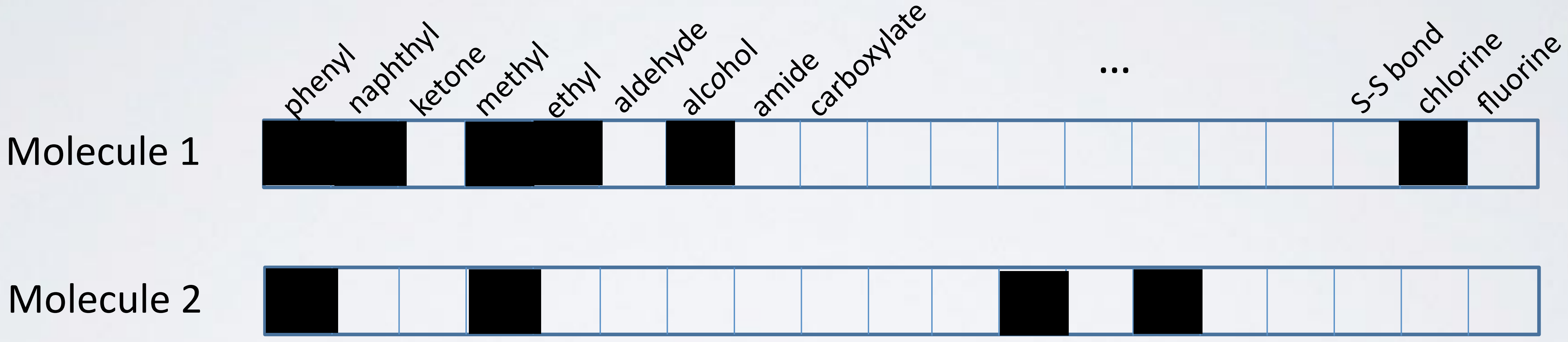


phenyl naphthyl ketone methyl ethyl aldehyde alcohol amide carboxylate … S-S bond chlorine fluorine

Molecule 1

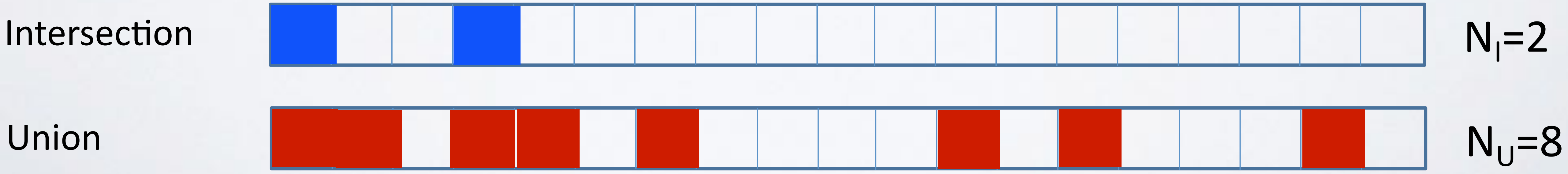
Molecule 2

CHEMICAL SIMILARITY FROM FINGERPRINTS



Tanimoto Similarity (or Jaccard Index), T

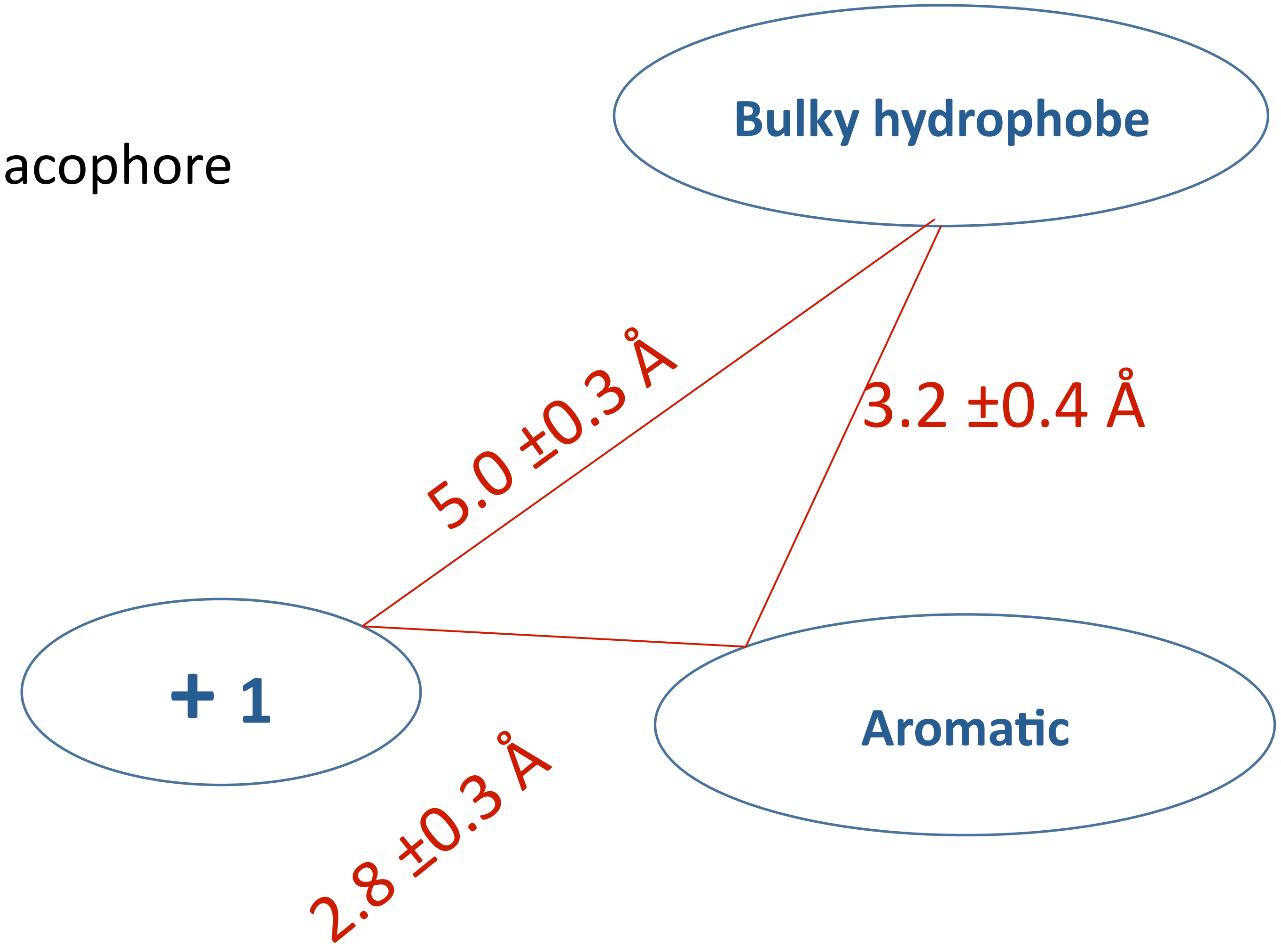
$$T \equiv \frac{N_I}{N_U} = 0.25$$



Pharmacophore Models

Φάρμακο (drug) + Φορά (carry)

A 3-point pharmacophore

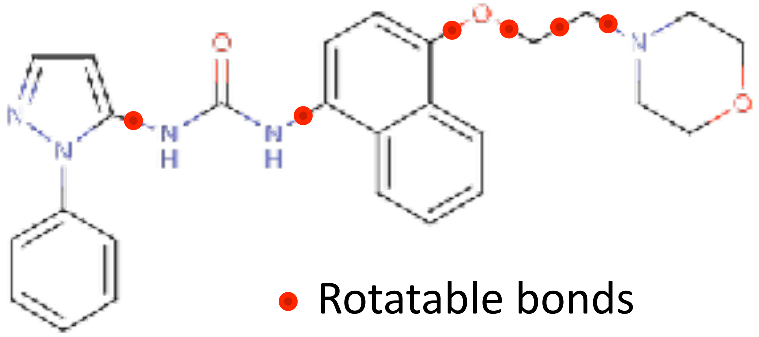


Molecular Descriptors

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

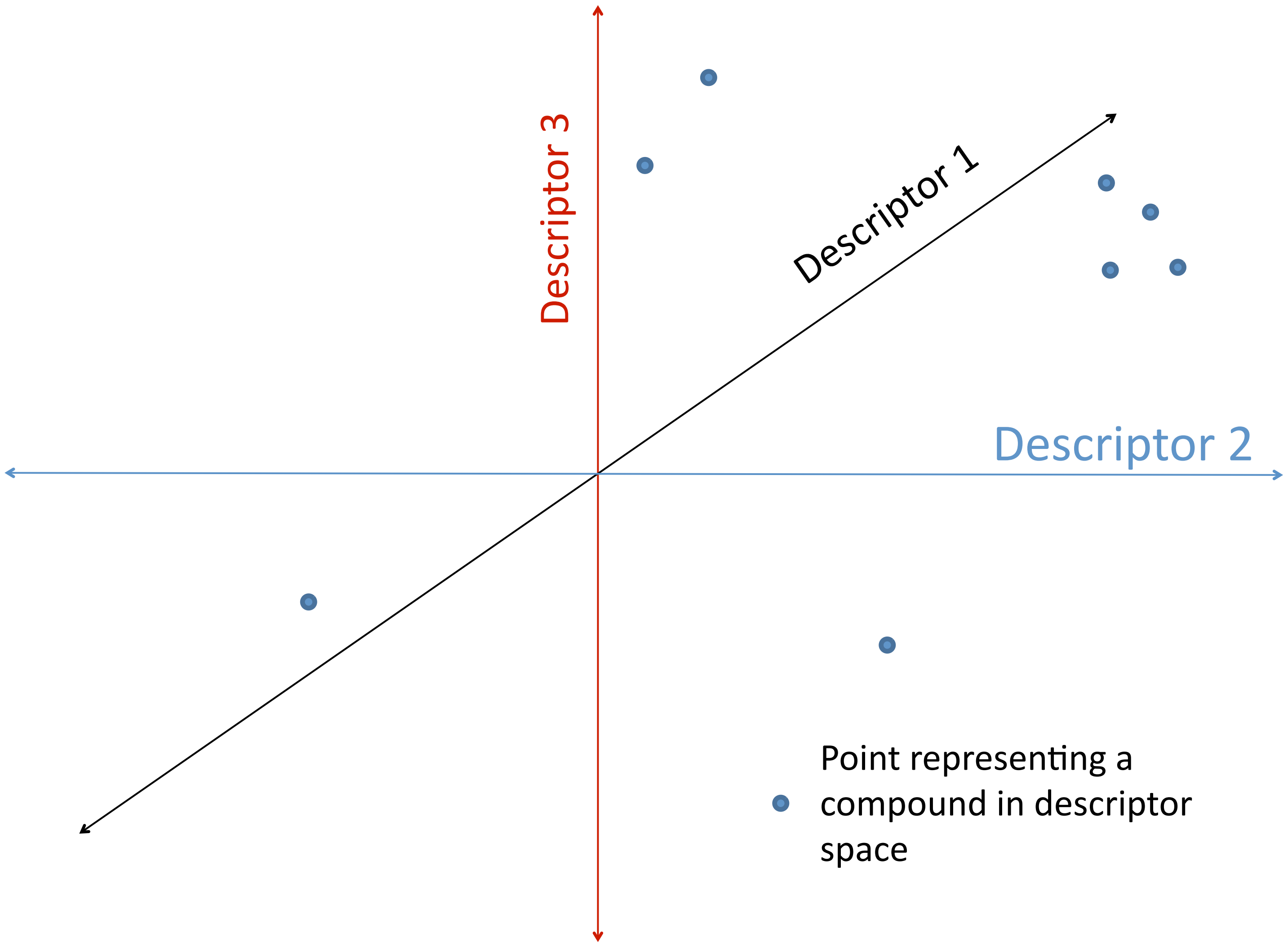
- branching index
- measures of linearity vs interconnectedness

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



Apply **multivariate statistics** and **machine learning** for descriptor-selection.
(e.g. partial least squares, support vector machines, random forest, etc.)

CAUTIONARY NOTES

- **“Everything should be made as simple as it can be but not simpler”**

  A model is **never perfect**. A model that is not quantitatively accurate in every respect does not preclude one from establishing results relevant to our understanding of biomolecules as long as the biophysics of the model are properly understood and explored.

- **Calibration of the parameters is an ongoing and imperfect process**

  Questions and hypotheses should always be designed such that they do not depend crucially on the precise numbers used for the various parameters.

- **A computational model is rarely universally right or wrong**

  A model may be accurate in some regards, inaccurate in others. These subtleties can only be uncovered by comparing to all available experimental data.

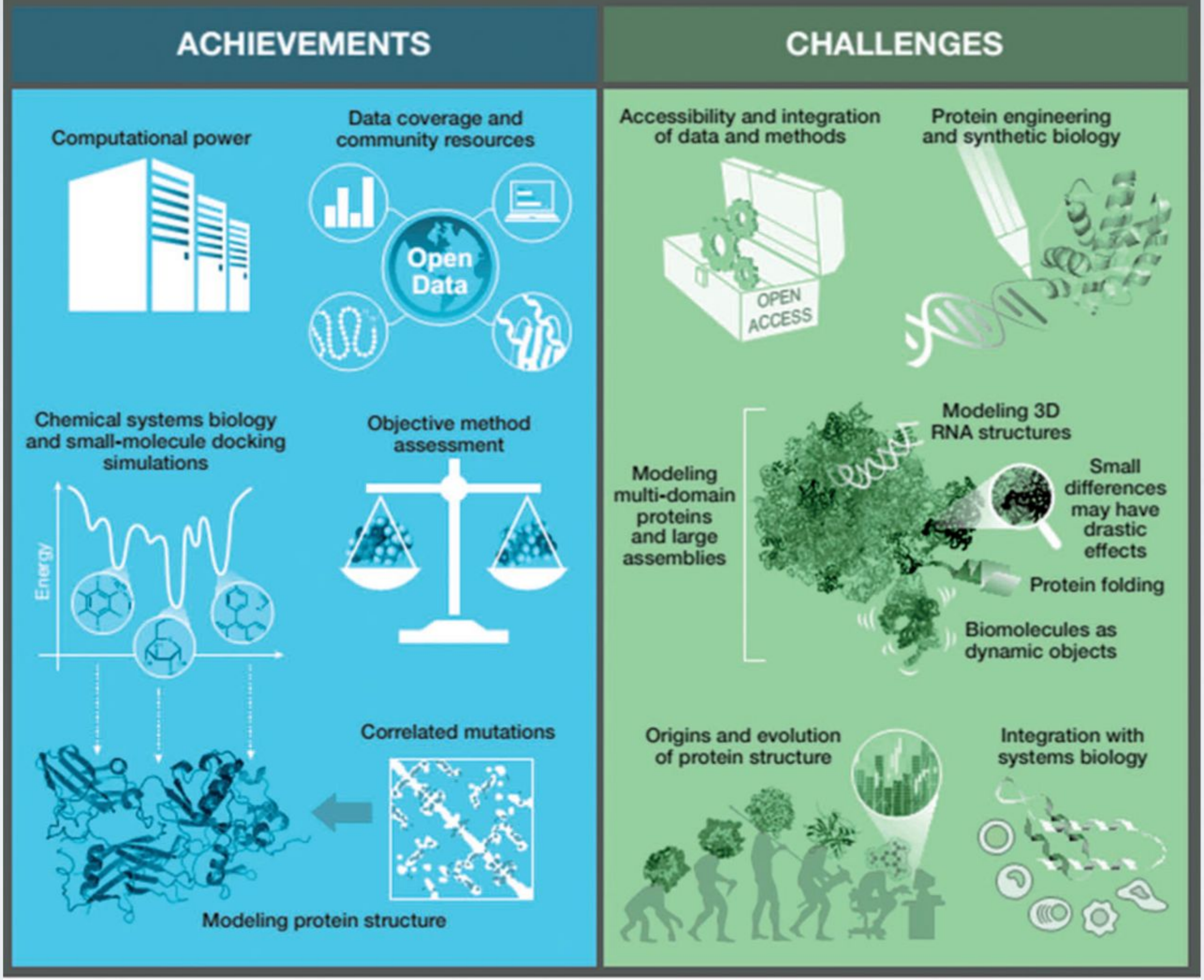
SUMMARY

- Structural bioinformatics is computer aided structural biology

- Described major motivations, goals and challenges of structural bioinformatics

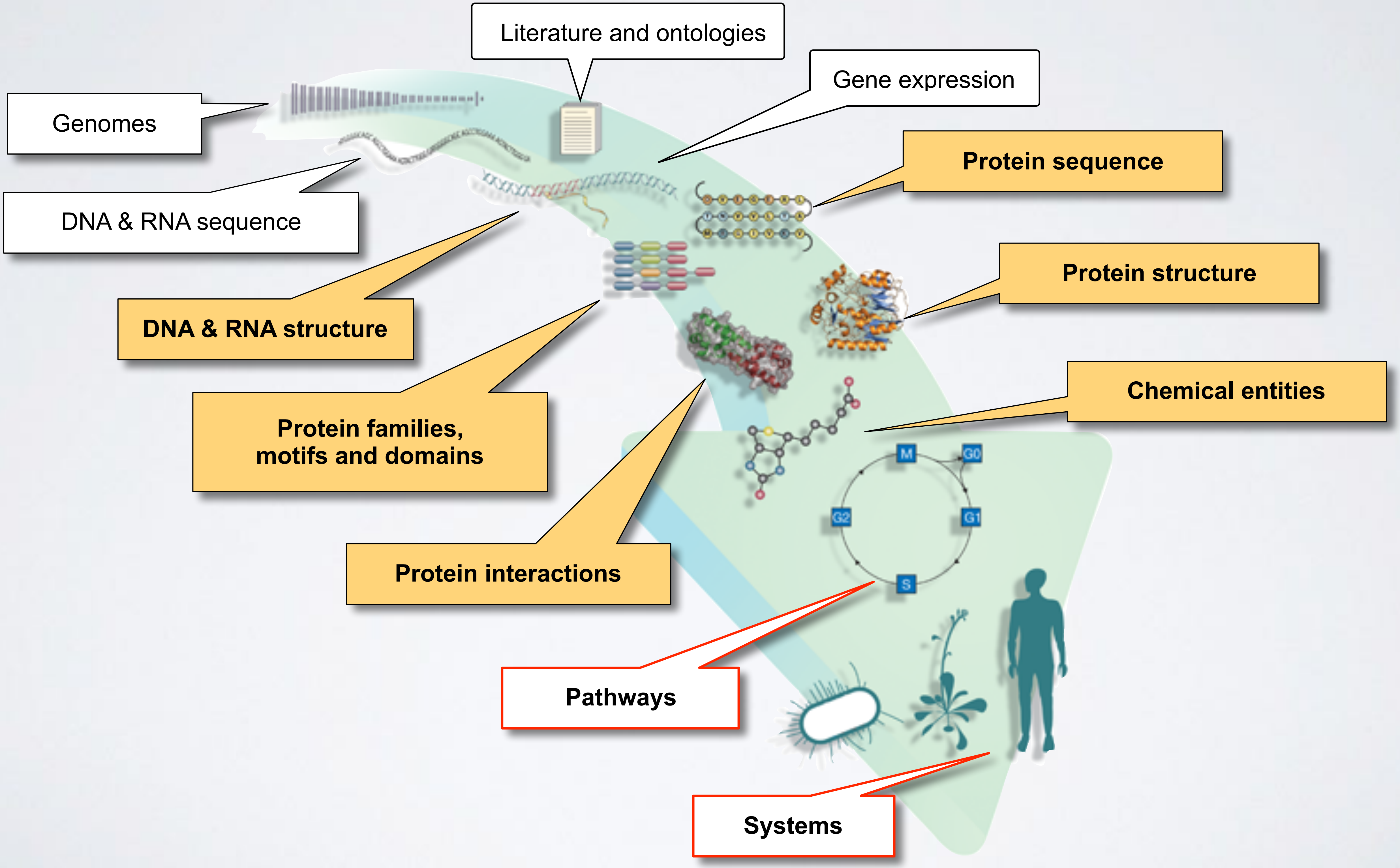
- Reviewed the fundamentals of protein structure

- Introduced both physics and knowledge based modeling approaches for describing the structure, energetics and dynamics of proteins computationally


ACHIEVEMENTS
CHALLENGES
Computational power
Data coverage and community resources
Open Data
Chemical systems biology and small-molecule docking simulations
Energy
Objective method assessment
Correlated mutations
Modeling protein structure
Accessibility and integration of data and methods
OPEN ACCESS
Protein engineering and synthetic biology
Modeling 3D RNA structures
Modeling multi-domain proteins and large assemblies
Small differences may have drastic effects
Protein folding
Biomolecules as dynamic objects
Origins and evolution of protein structure
Integration with systems biology

Ilan Samish et al. Bioinformatics 2015;31:146-150

INFORMING SYSTEMS BIOLOGY?



Literature and ontologies

Genomes

Gene expression

DNA & RNA sequence

Protein sequence

DNA & RNA structure

Protein structure

Protein families, motifs and domains

Chemical entities

Protein interactions

Pathways

Systems