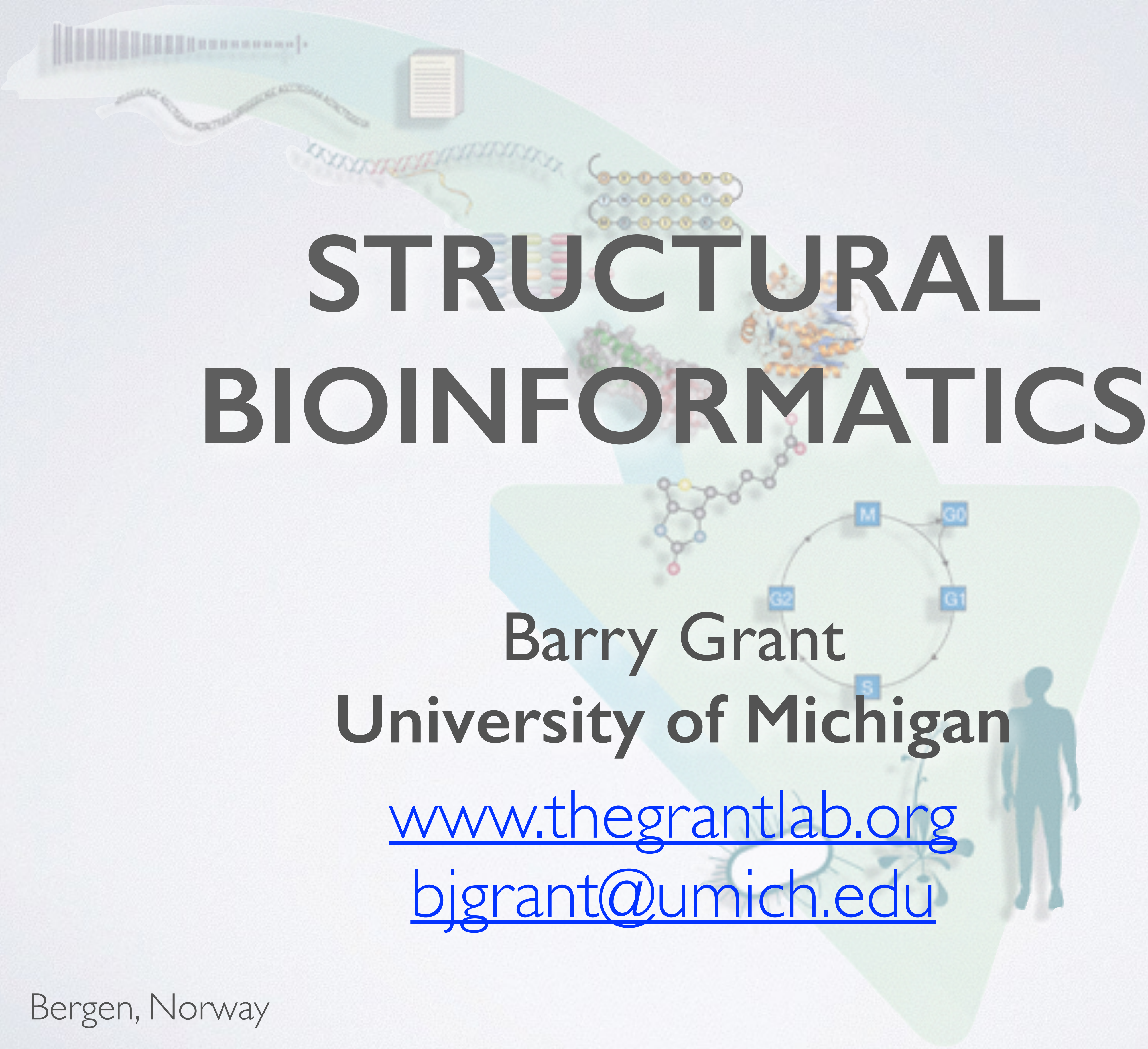


STRUCTURAL BIOINFORMATICS

Barry Grant
University of Michigan

www.thegrantlab.org
bjgrant@umich.edu

Bergen, Norway

28-Sep-2015

Objective:

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

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 the application of computers to the collection, archiving, organization, and interpretation of biological data.*”

[Orengo, 2003]

… A hybrid of biology and computer science

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

[Orengo, 2003]

Bioinformatics is computer aided biology!

So what is **structural bioinformatics**?

So what is **structural bioinformatics**?

… computer aided structural biology!

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

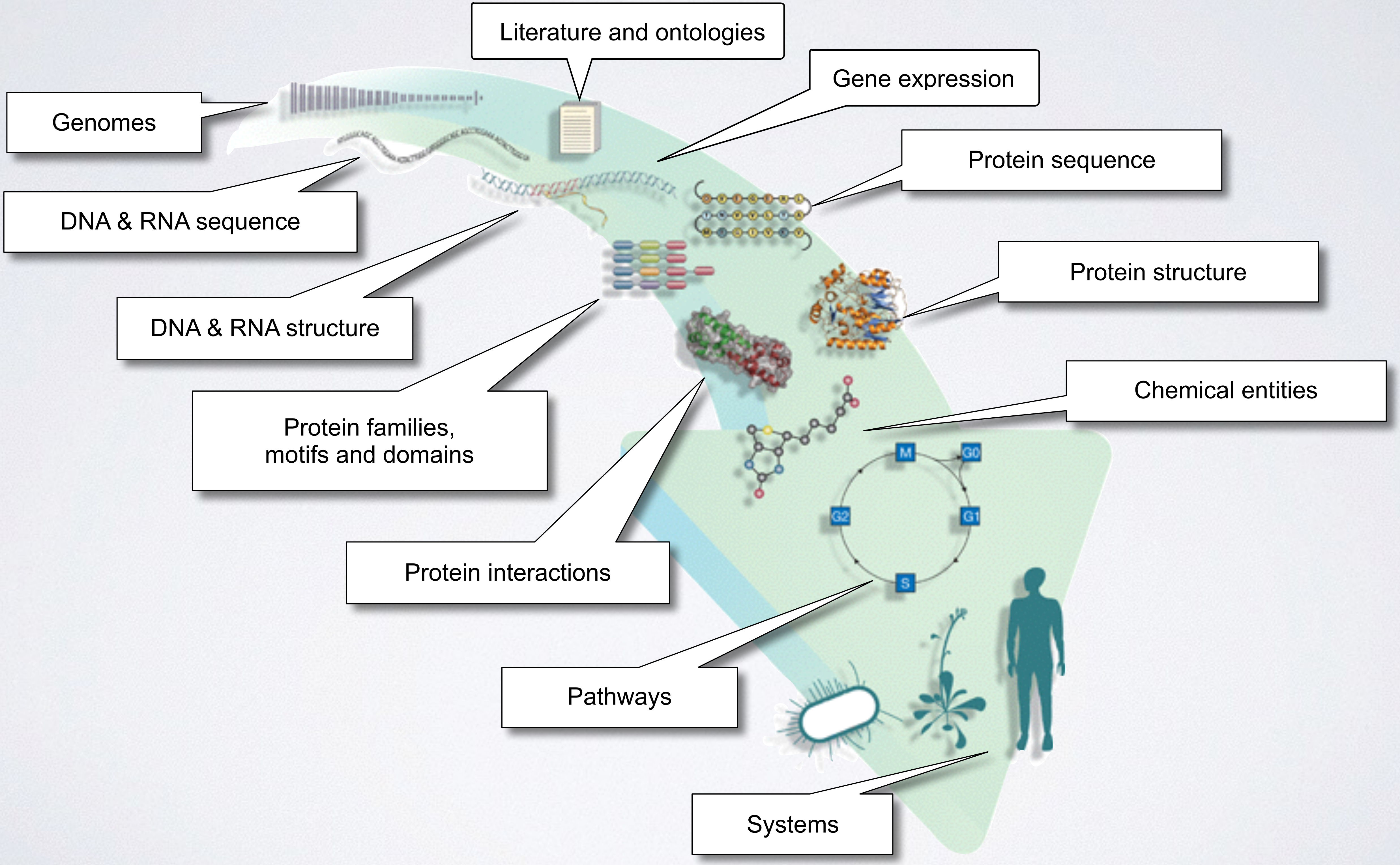
Why should we care?

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

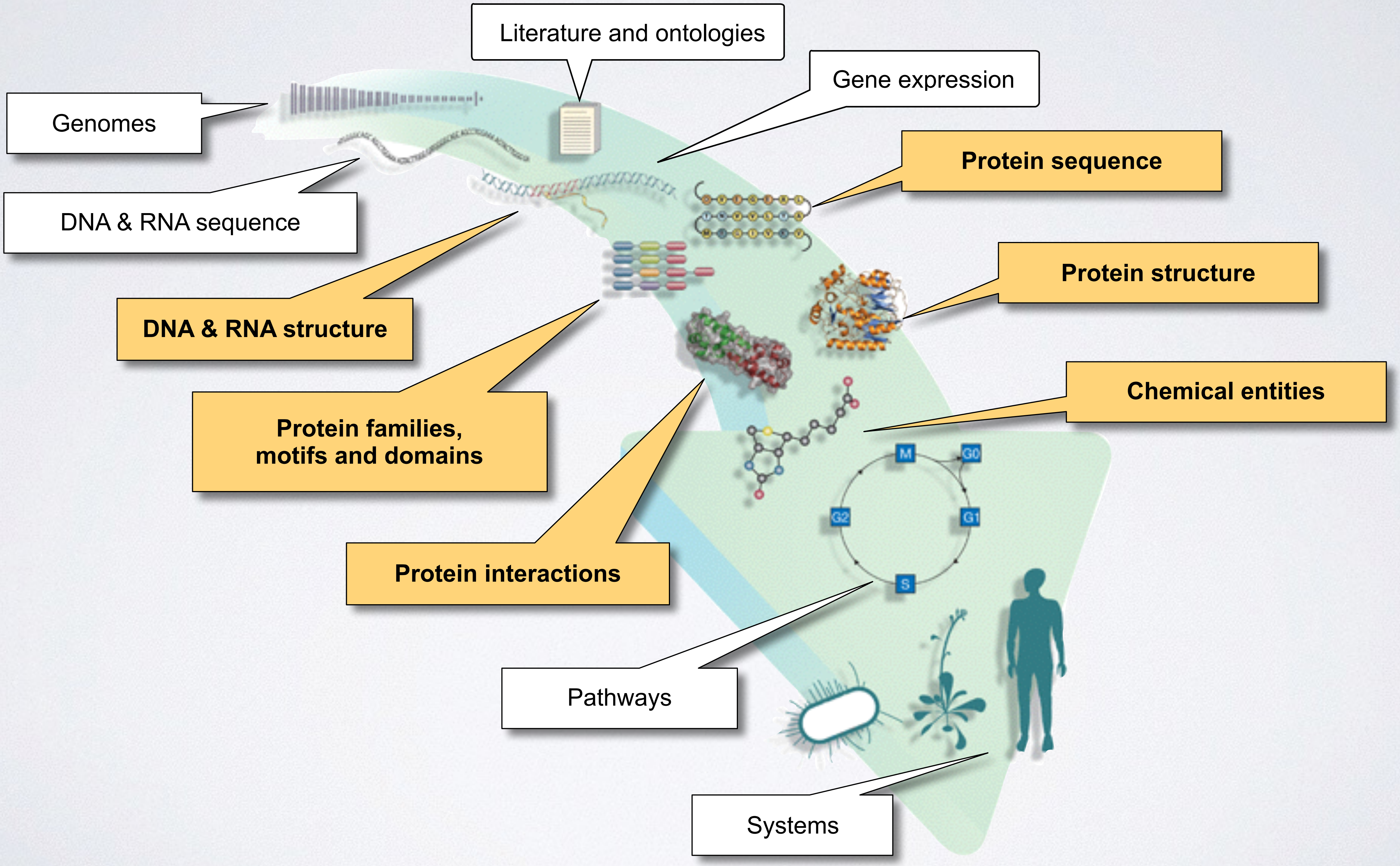
Protein families, motifs and domains

Protein interactions

Pathways

Systems

STRUCTURAL DATA IS CENTRAL



Literature and ontologies

Genomes

Gene expression

Protein sequence

DNA & RNA sequence

Protein structure

DNA & RNA structure

Chemical entities

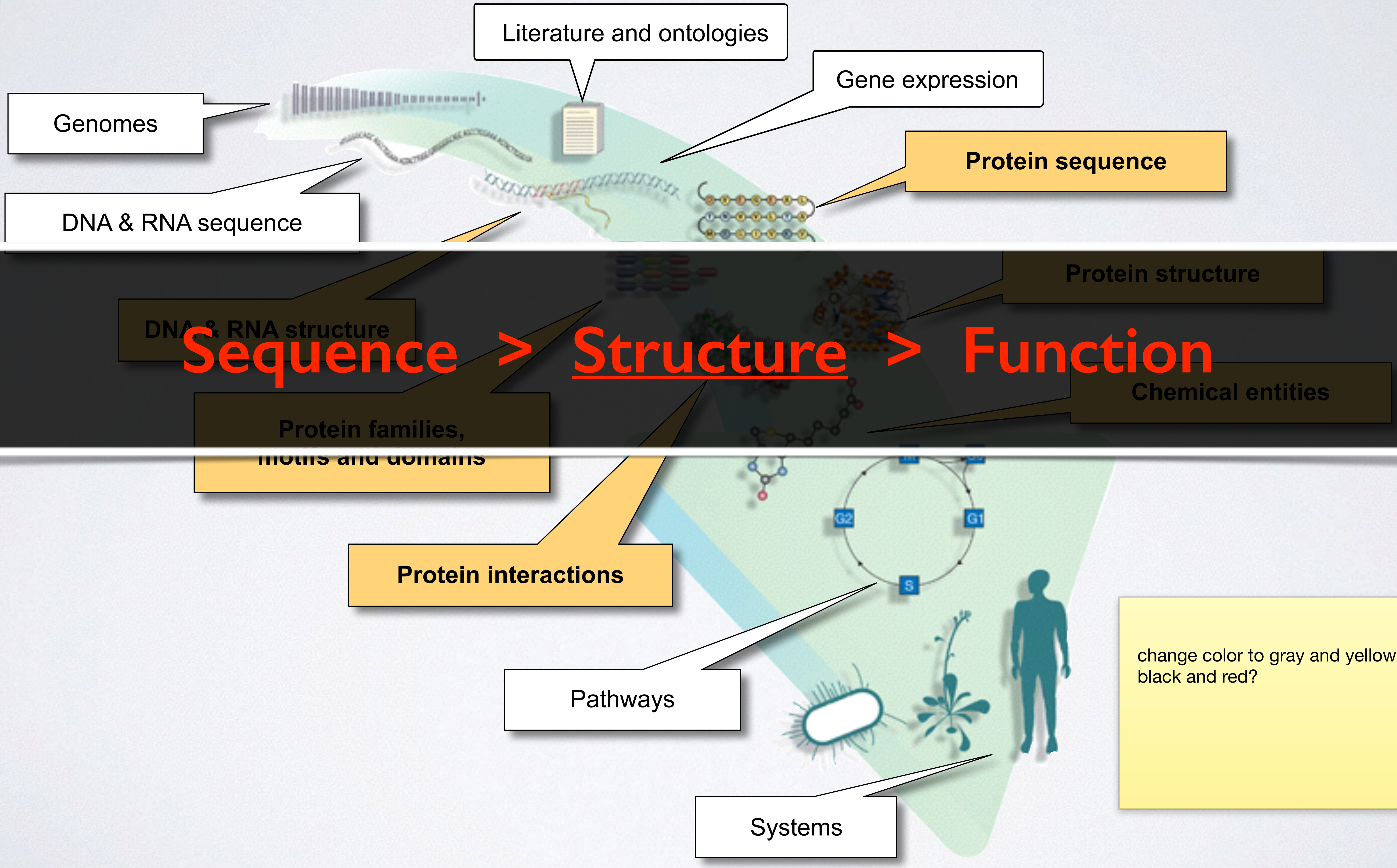
Protein families, motifs and domains

Protein interactions

Pathways

Systems

STRUCTURAL DATA IS CENTRAL



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

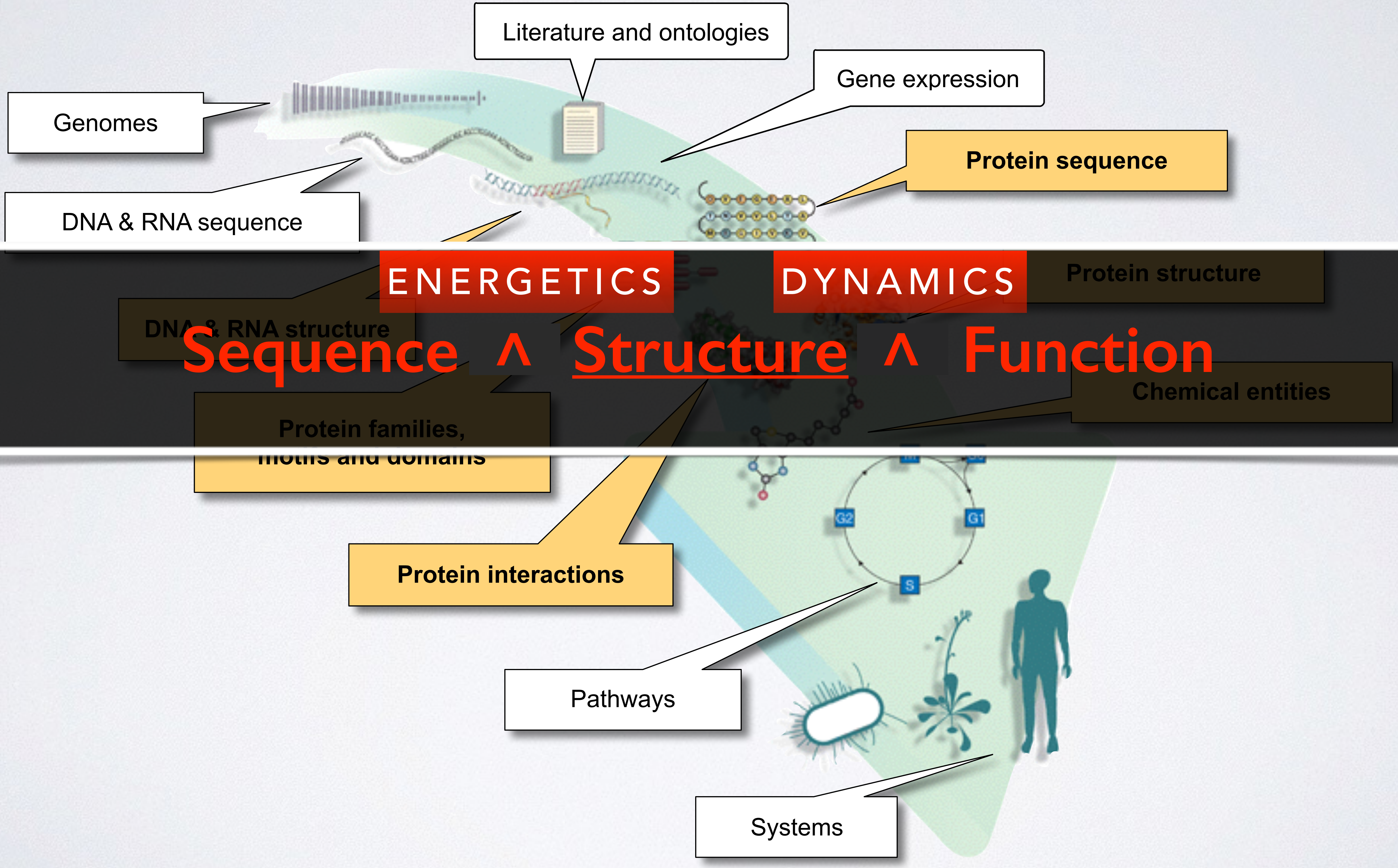
Pathways

Systems

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

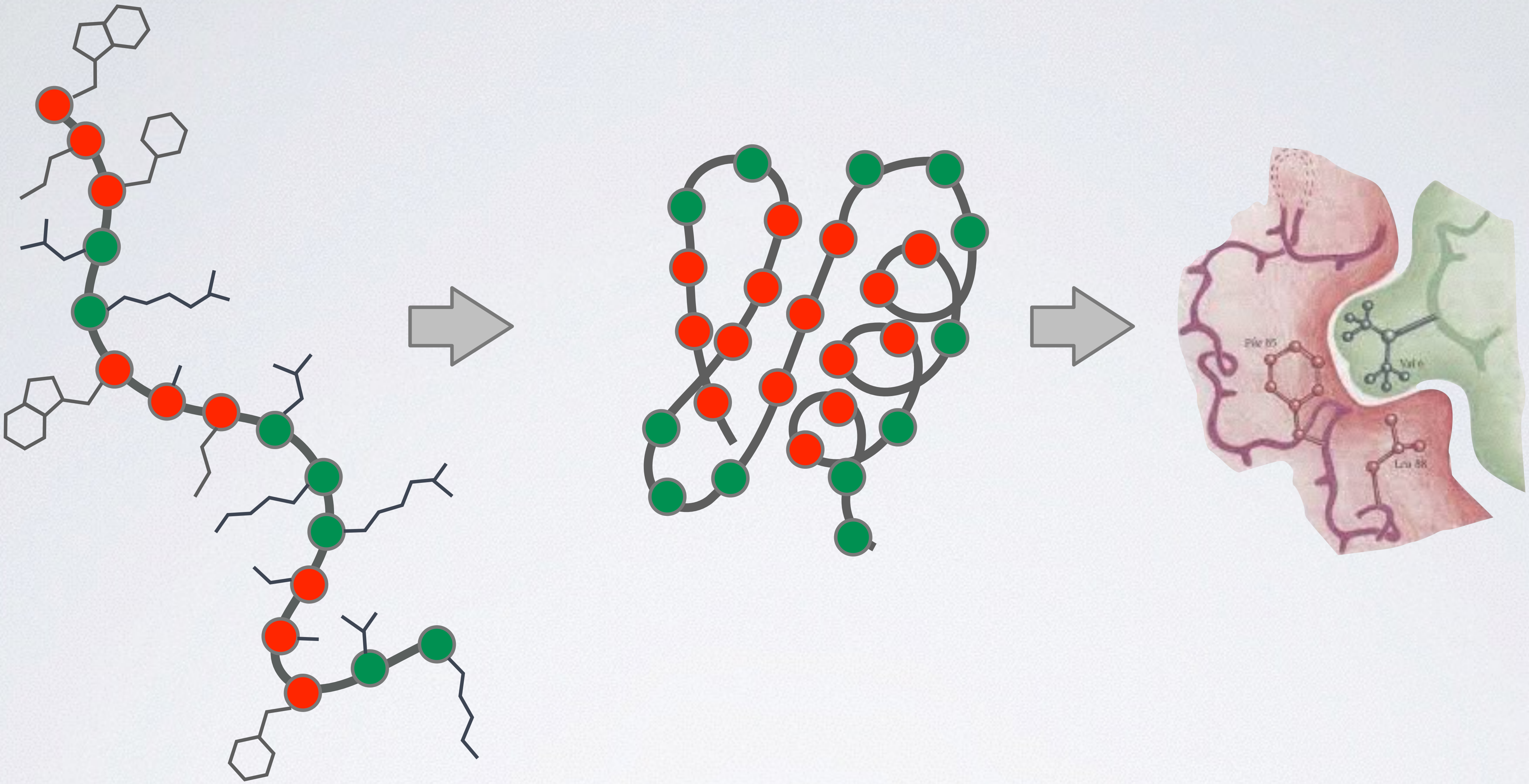
Protein interactions

Pathways

Systems

ENERGETICS DYNAMICS

Sequence ∧ Structure ∧ Function



Sequence

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

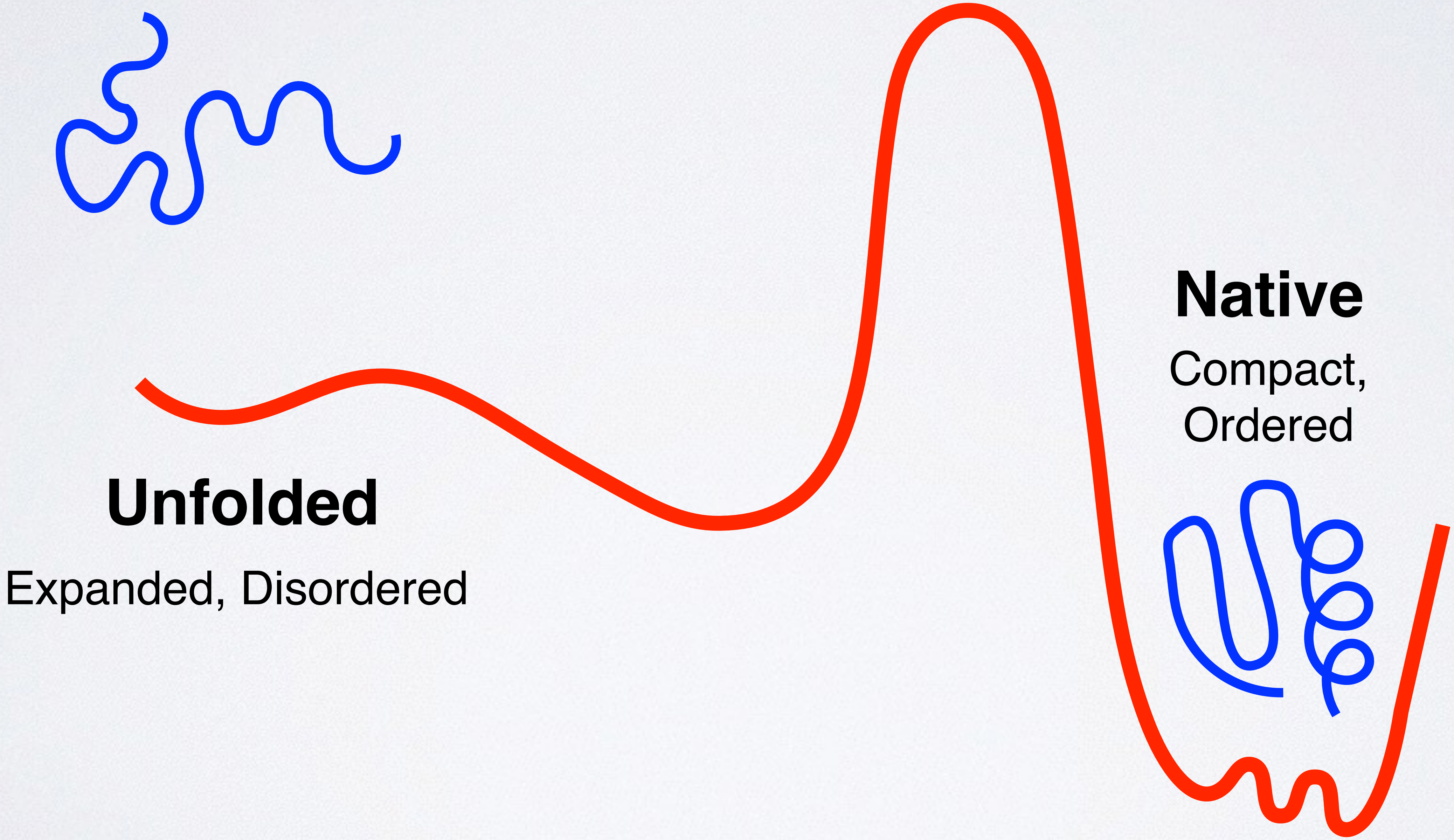
Structure

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

Function

- Active in specific “conformations”
- Specific associations & precise reactions

KEY CONCEPT: ENERGY LANDSCAPE



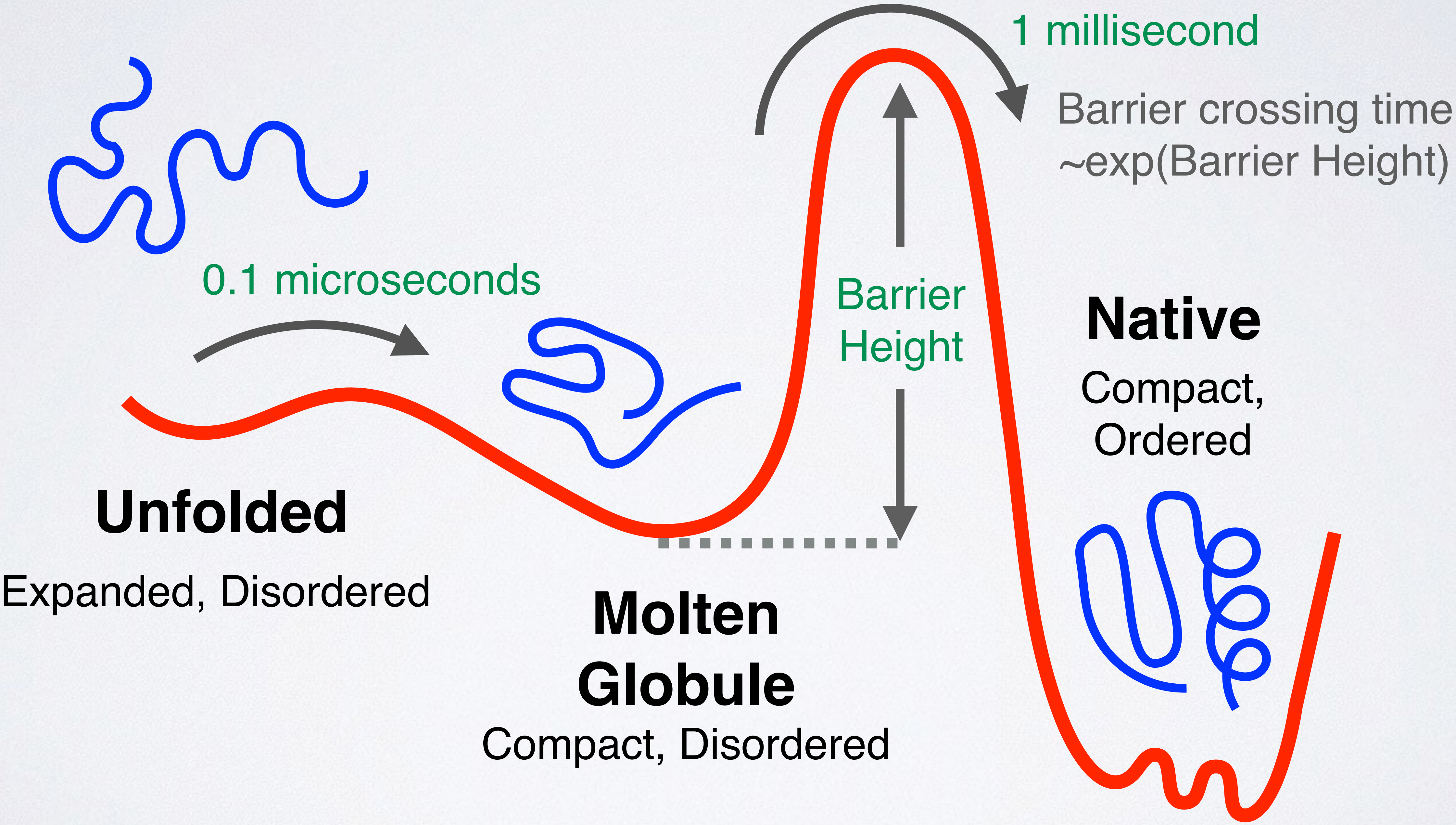
Unfolded

Expanded, Disordered

Native

Compact,
Ordered

KEY CONCEPT: ENERGY LANDSCAPE



1 millisecond

Barrier crossing time ~exp(Barrier Height)

0.1 microseconds

Barrier Height

Native

Compact, Ordered

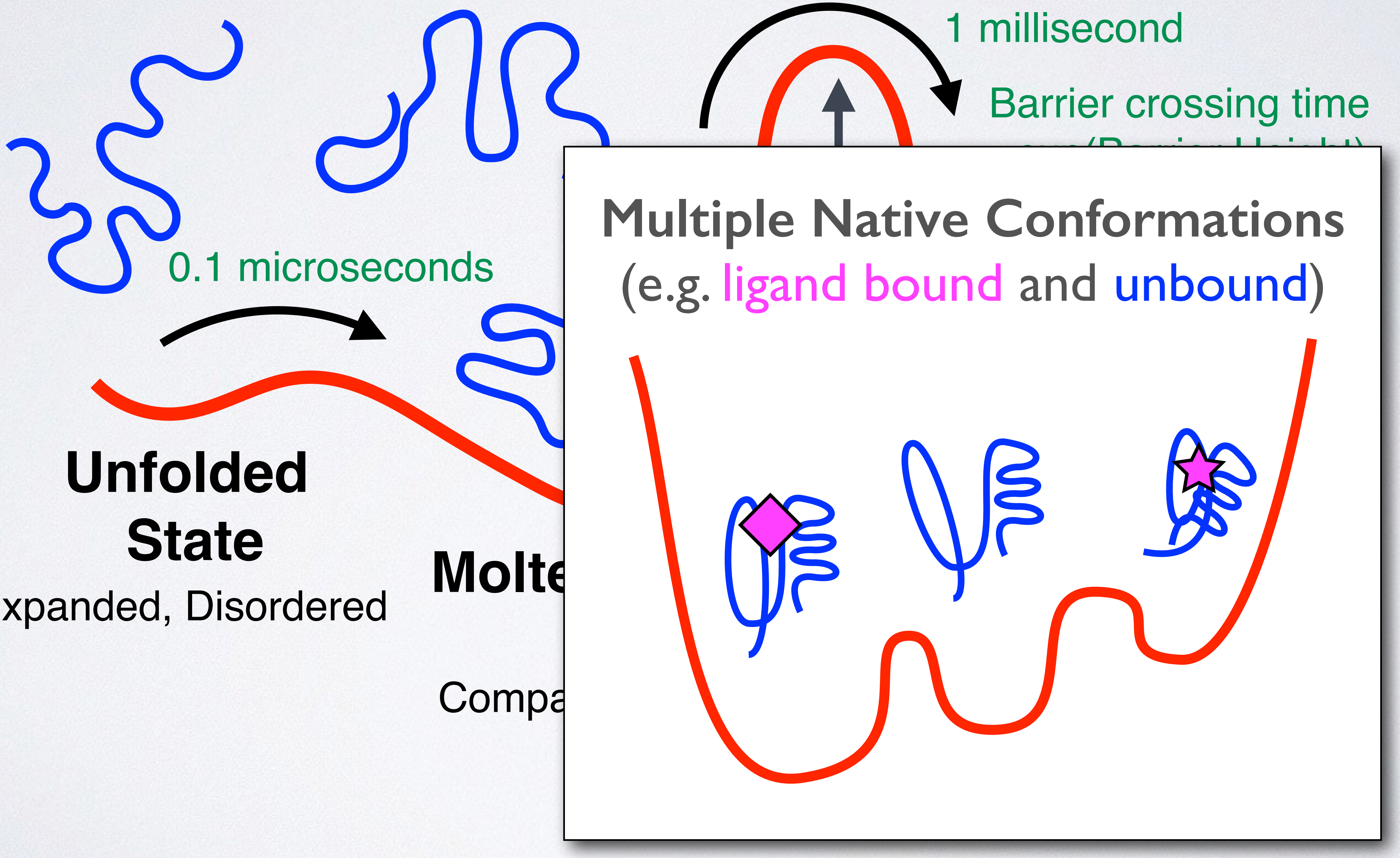
Unfolded

Expanded, Disordered

Molten Globule

Compact, Disordered

KEY CONCEPT: **ENERGY LANDSCAPE**



1 millisecond

Barrier crossing time

0.1 microseconds

Unfolded State

Expanded, Disordered

Molte

Compa

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

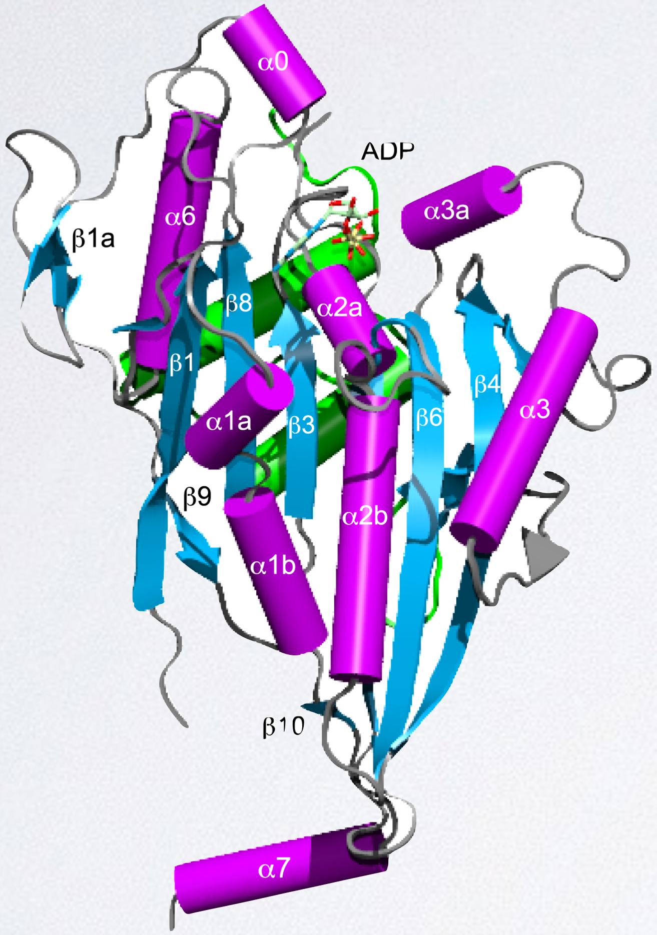
OUTLINE:

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

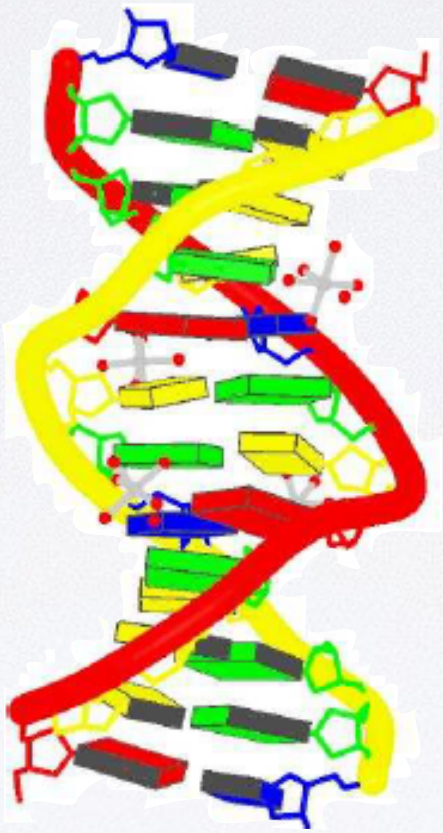
- **Fundamentals of protein structure**
  - Composition, form, forces and dynamics

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

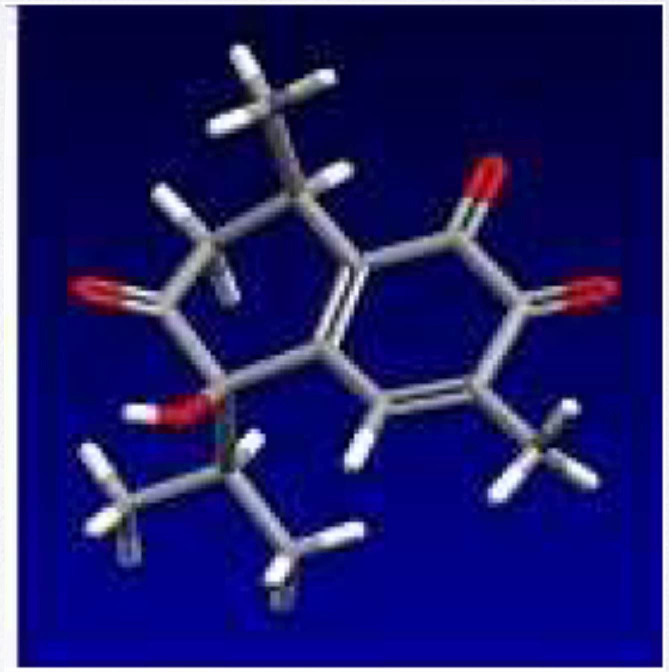
TRADITIONAL FOCUS **PROTEIN**, **DNA** AND **SMALL MOLECULE** DATA SETS WITH **MOLECULAR STRUCTURE**



Protein
(PDB)



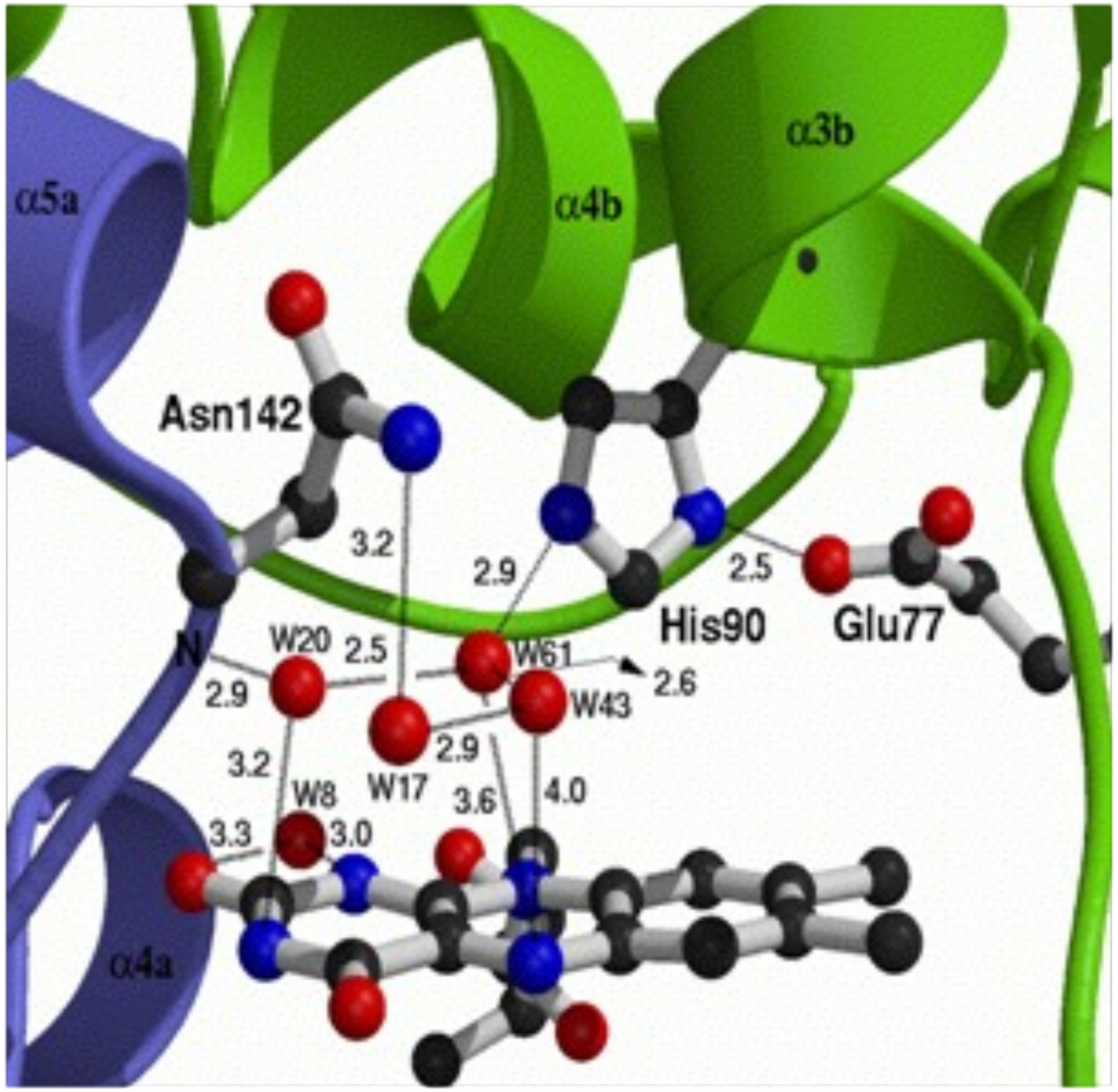
DNA
(NDB)



Small Molecules
(CCDB)

Motivation 1:
Detailed understanding of molecular interactions

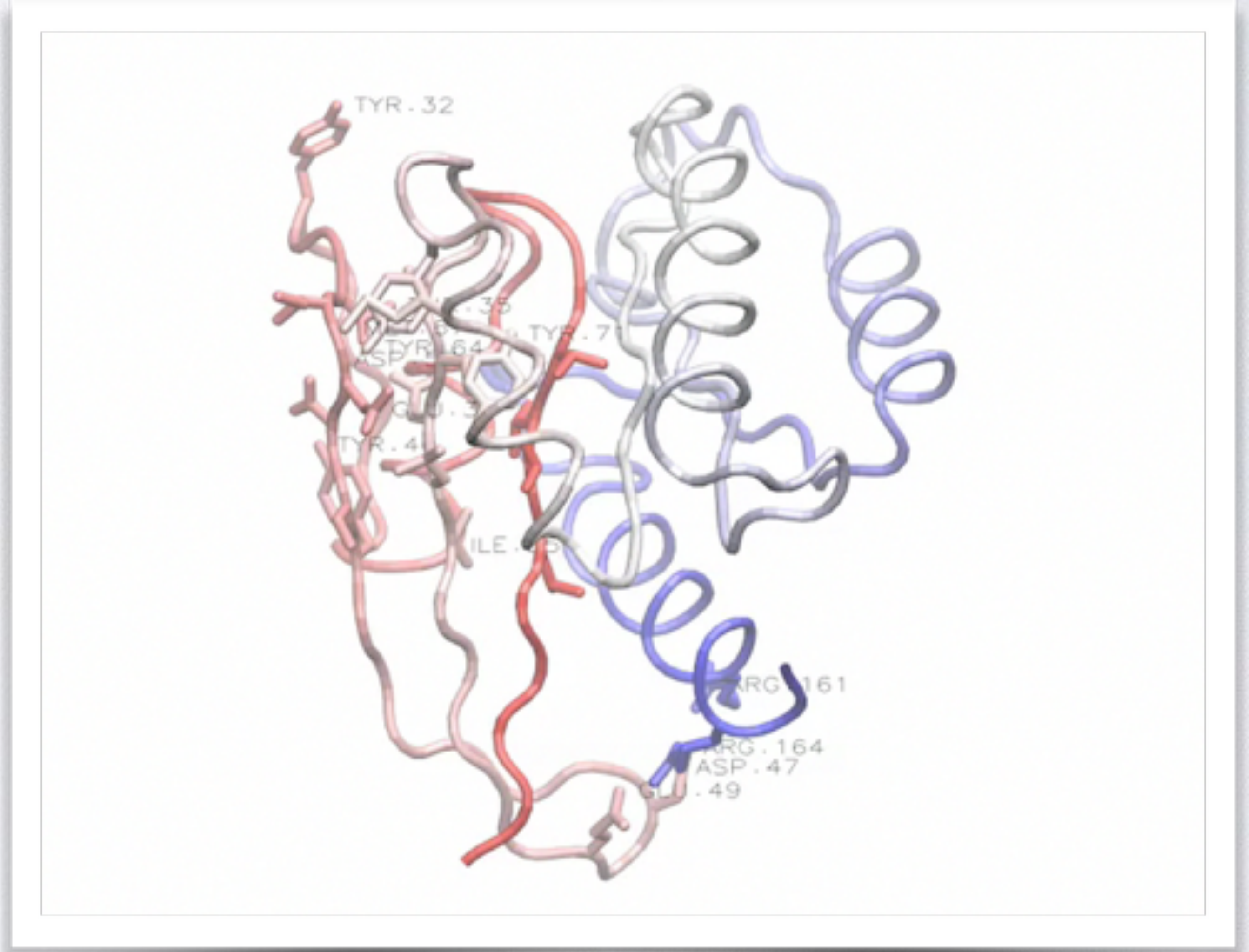
Provides an invaluable structural context for conservation and mechanistic analysis leading to functional insight.



Motivation 1:

Detailed understanding of molecular interactions

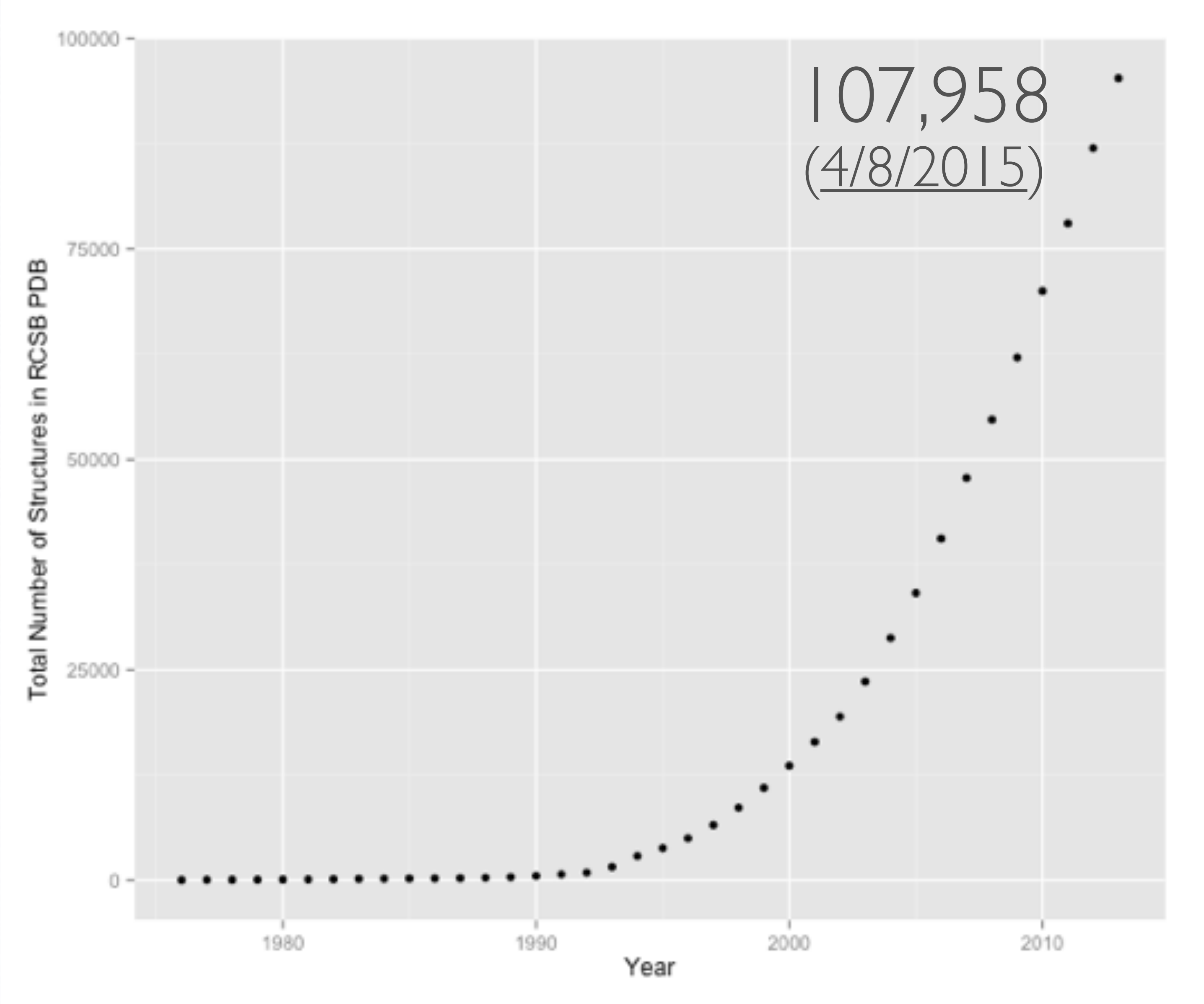
Computational modeling can provide detailed insight into functional interactions, their regulation and potential consequences of perturbation.



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

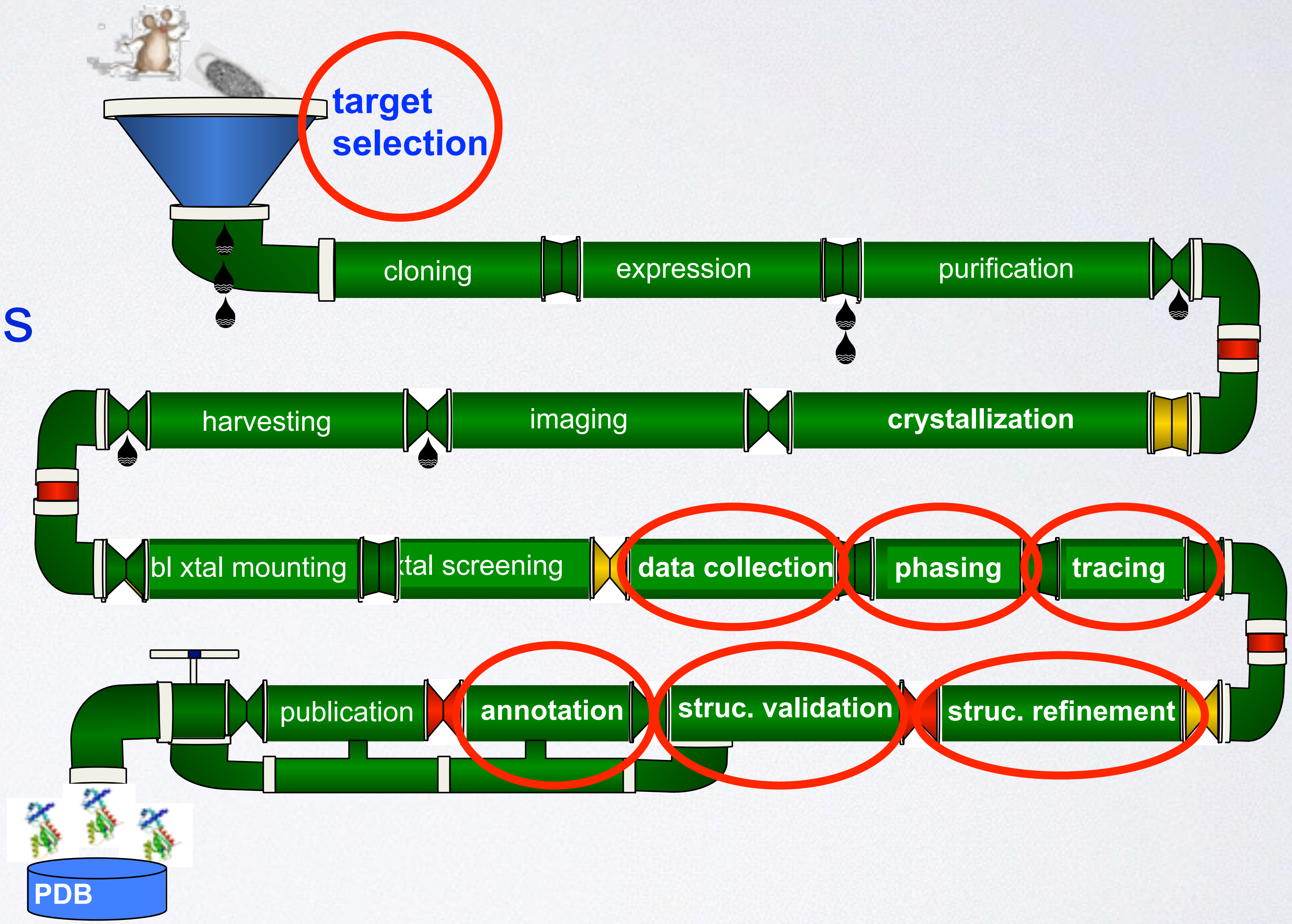


107,958
(4/8/2015)

Data from: http://www.rcsb.org/pdb/statistics/

Motivation 2:

Lots of structural data is becoming available

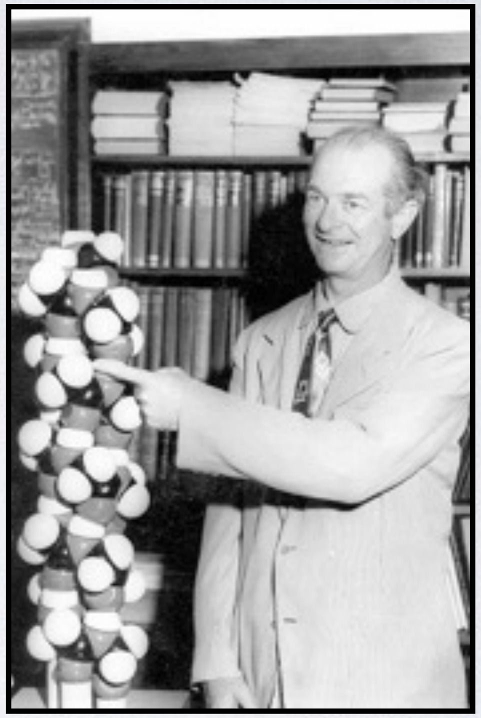


Structural Genomics has contributed to driving down the cost and time required for structural determination

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

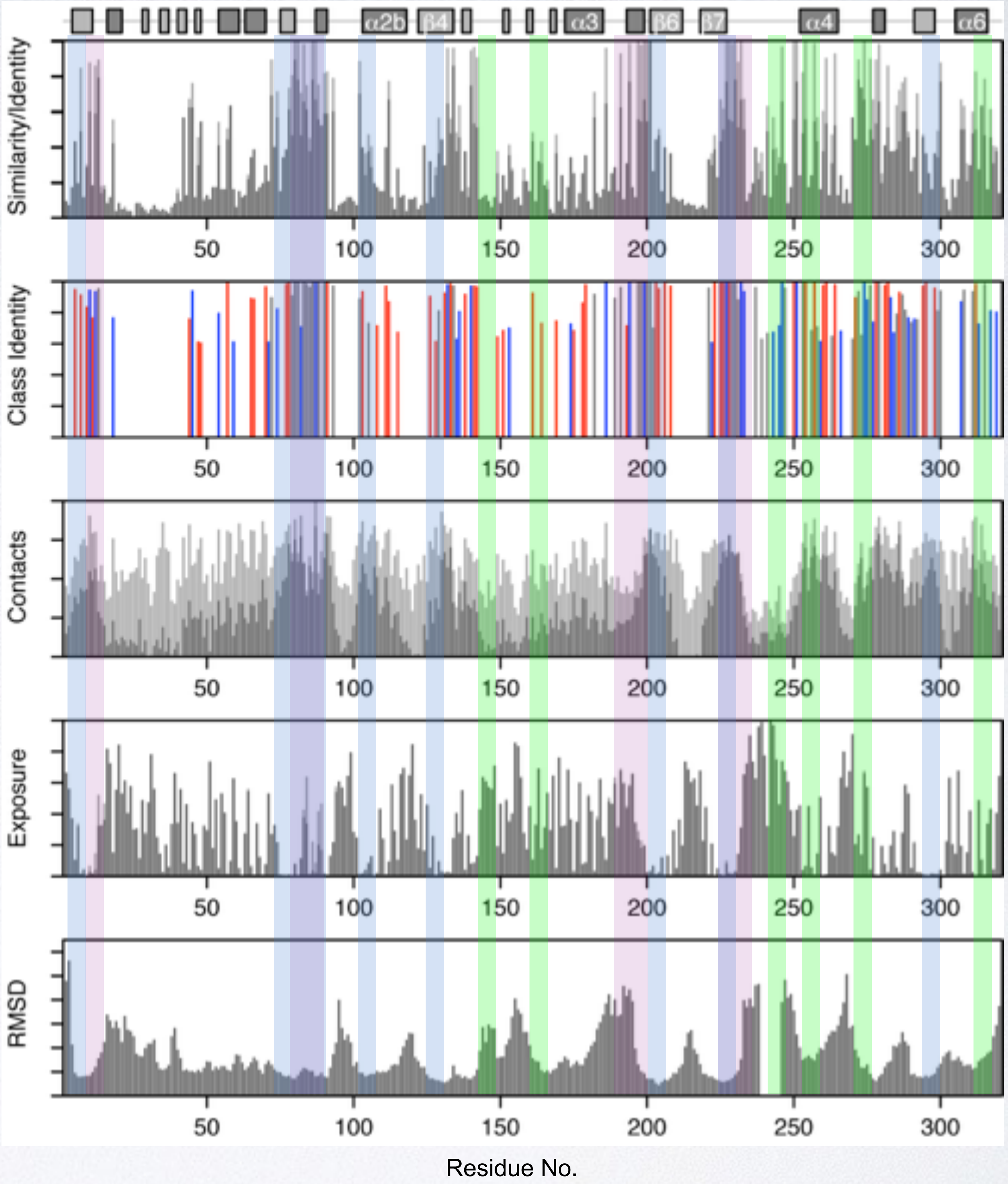
- Structure allows identification of more distant evolutionary relationships

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)

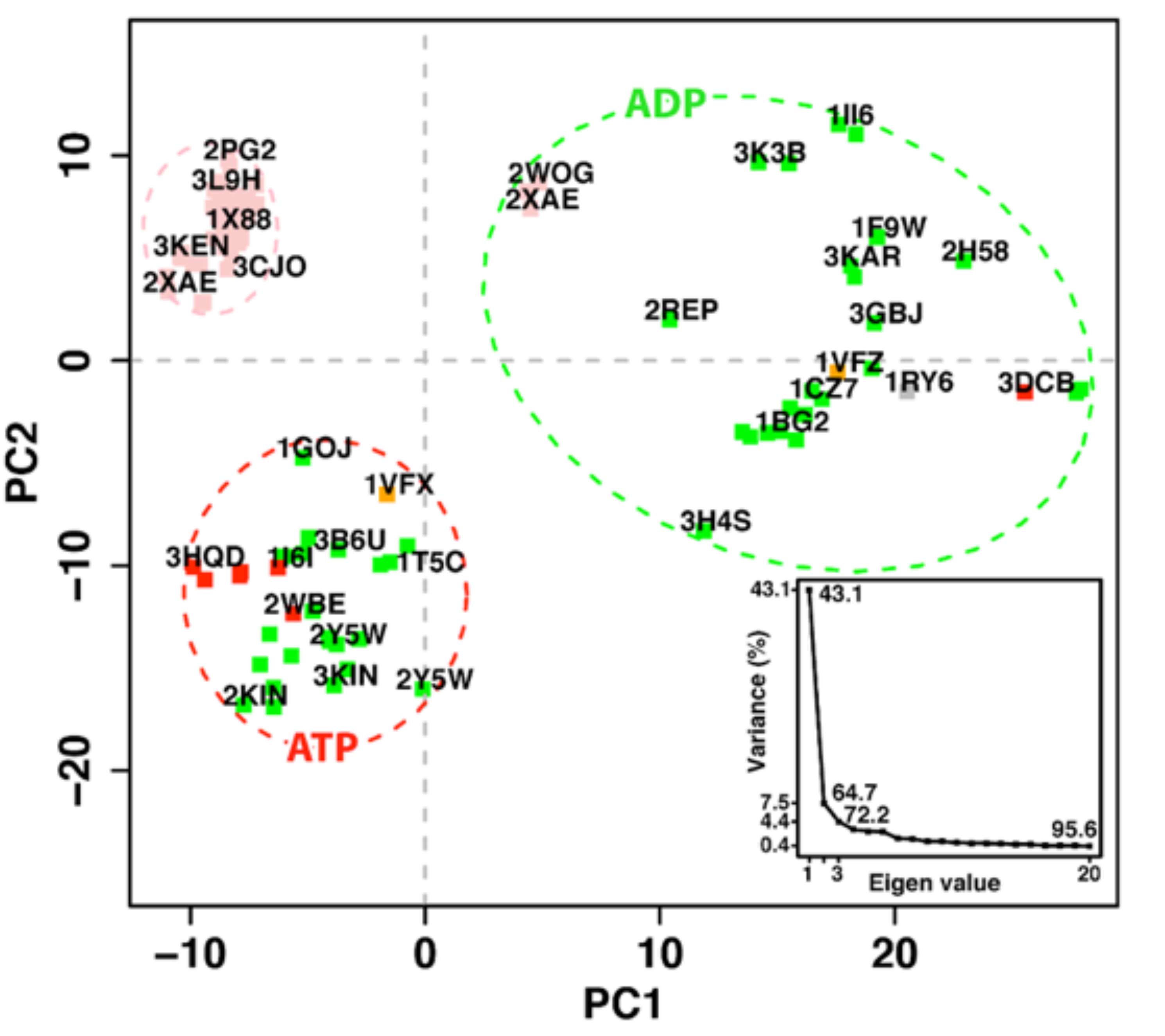
Goals:

- Analysis
- Visualization
- Comparison
- Prediction
- Design

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

Goals:

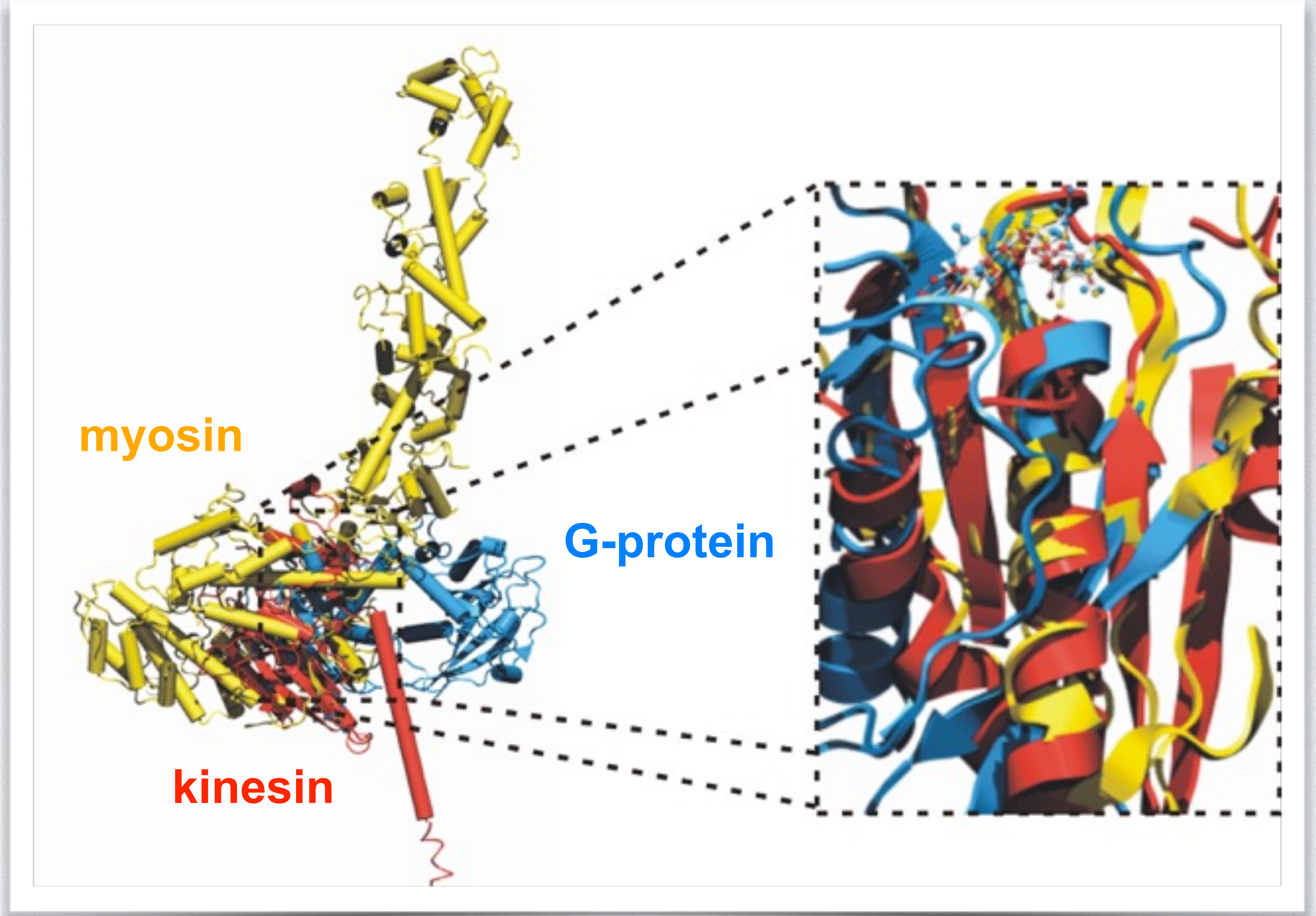
- Analysis
- Visualization
- Comparison
- Prediction
- Design



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

Goals:

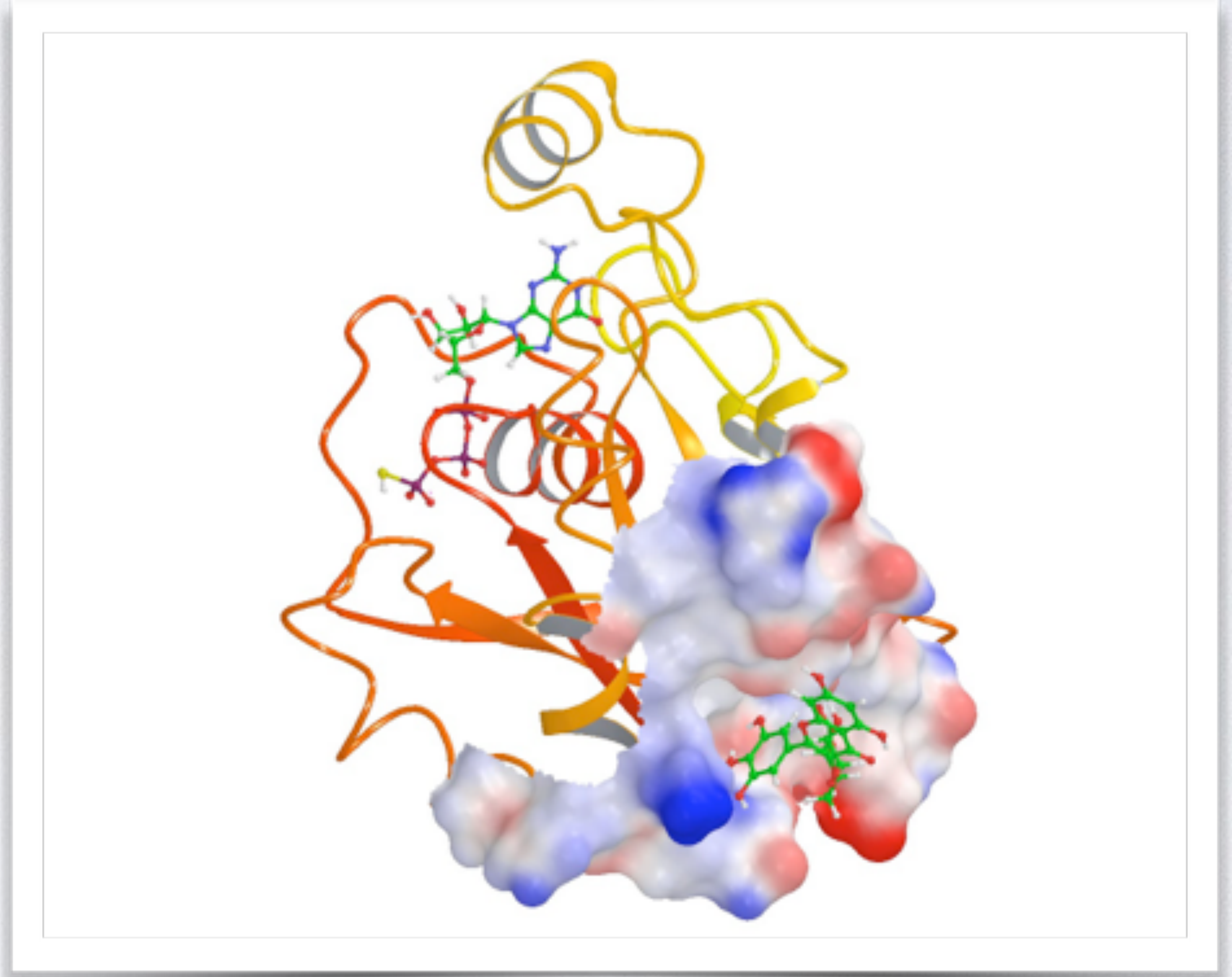
- Analysis
- Visualization
- Comparison
- Prediction
- Design



Grant *et al.* unpublished

Goals:

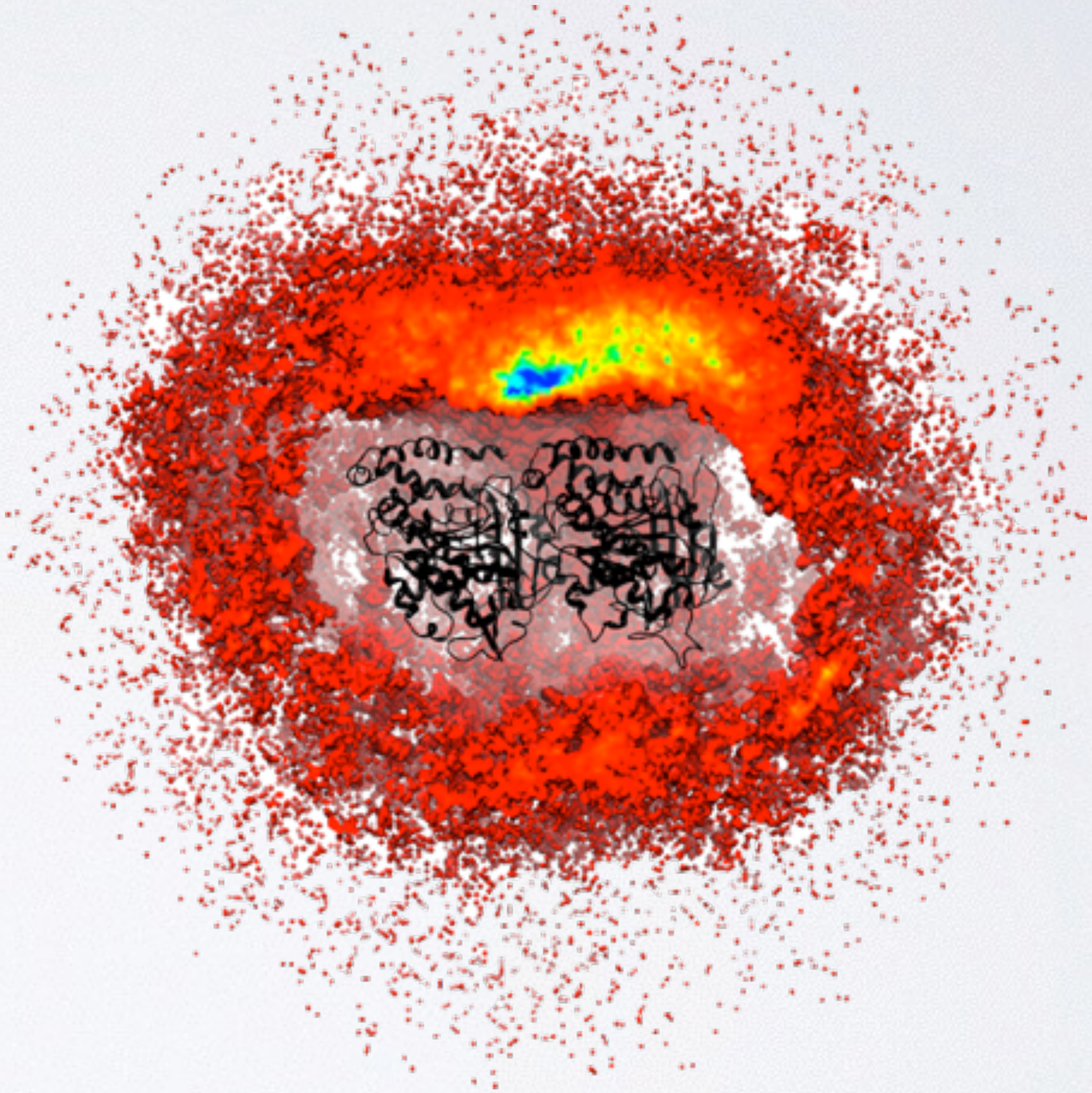
- Analysis
- Visualization
- Comparison
- Prediction
- Design



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...

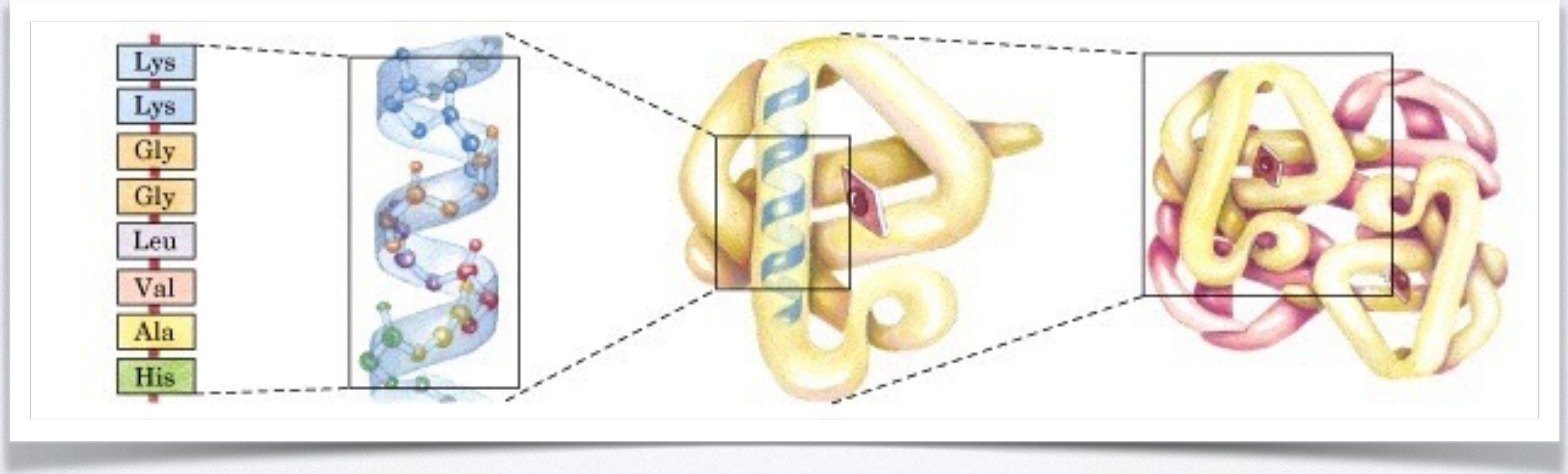
With applications to Biology, Medicine, Agriculture and Industry

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

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

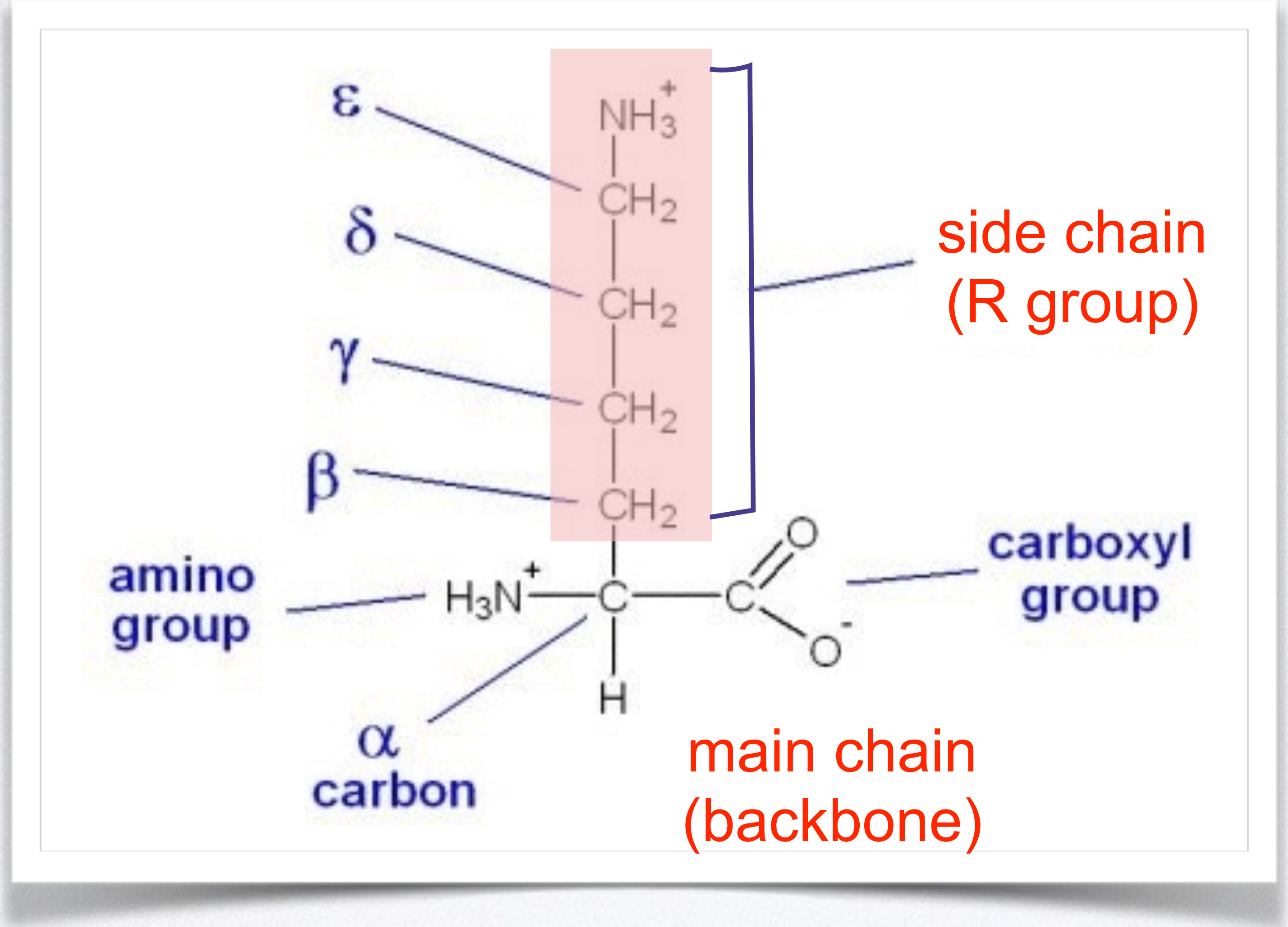
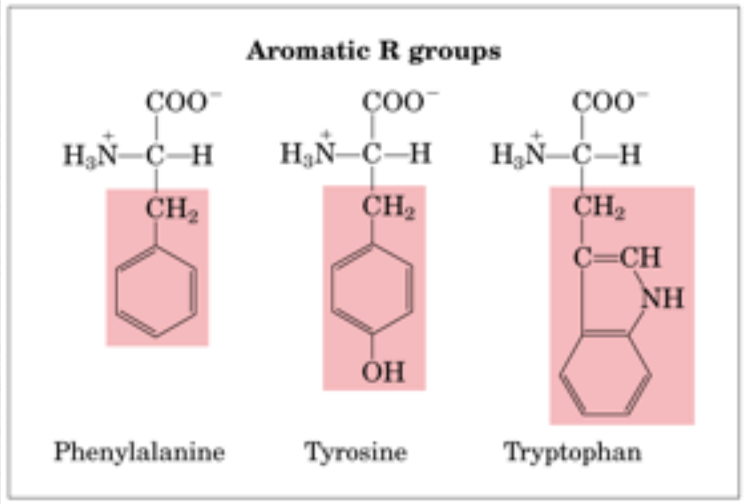
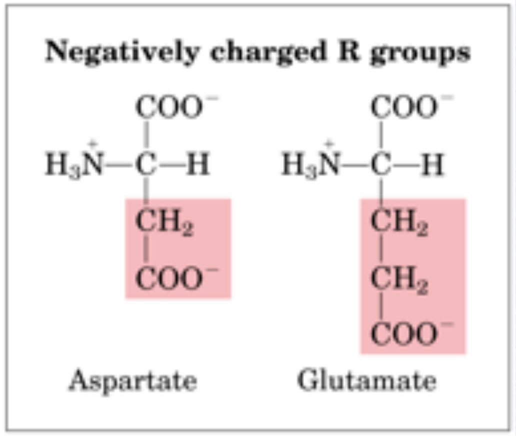
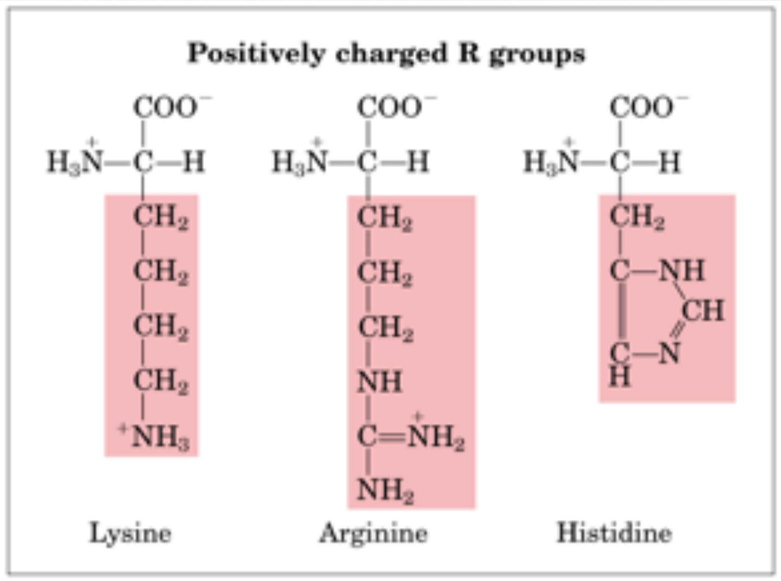


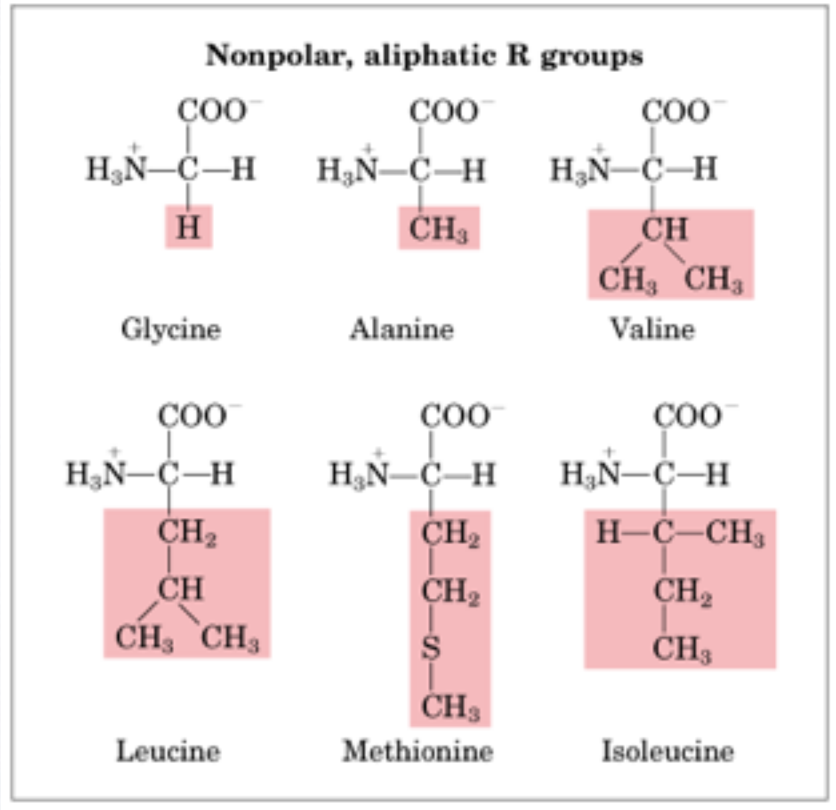
Image from: http://www.ncbi.nlm.nih.gov/books/NBK21581/

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


Aromatic R groups


Negatively charged R groups


Positively charged R groups


Nonpolar, aliphatic R groups

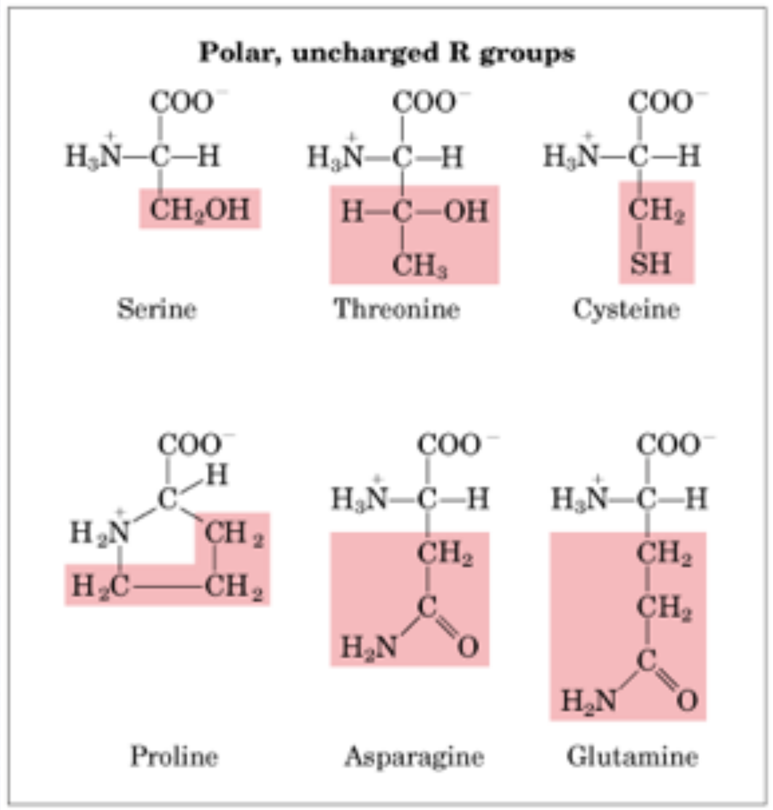
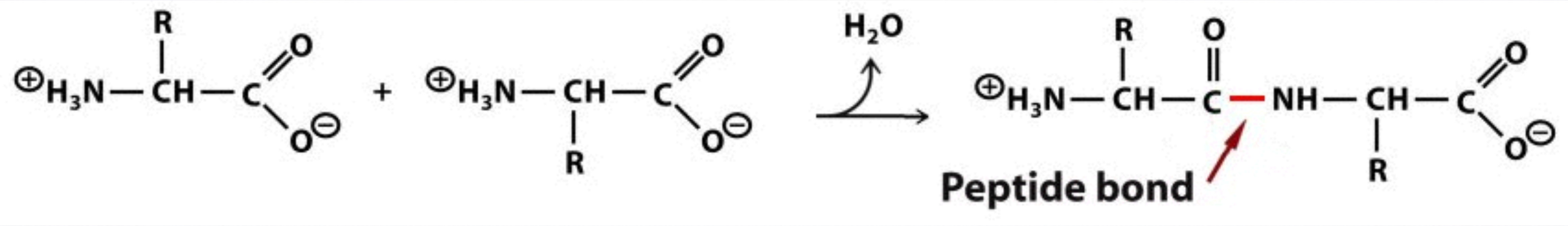

Polar, uncharged R groups

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

AMINO ACIDS POLYMERIZE THROUGH **PEPTIDE BOND** FORMATION



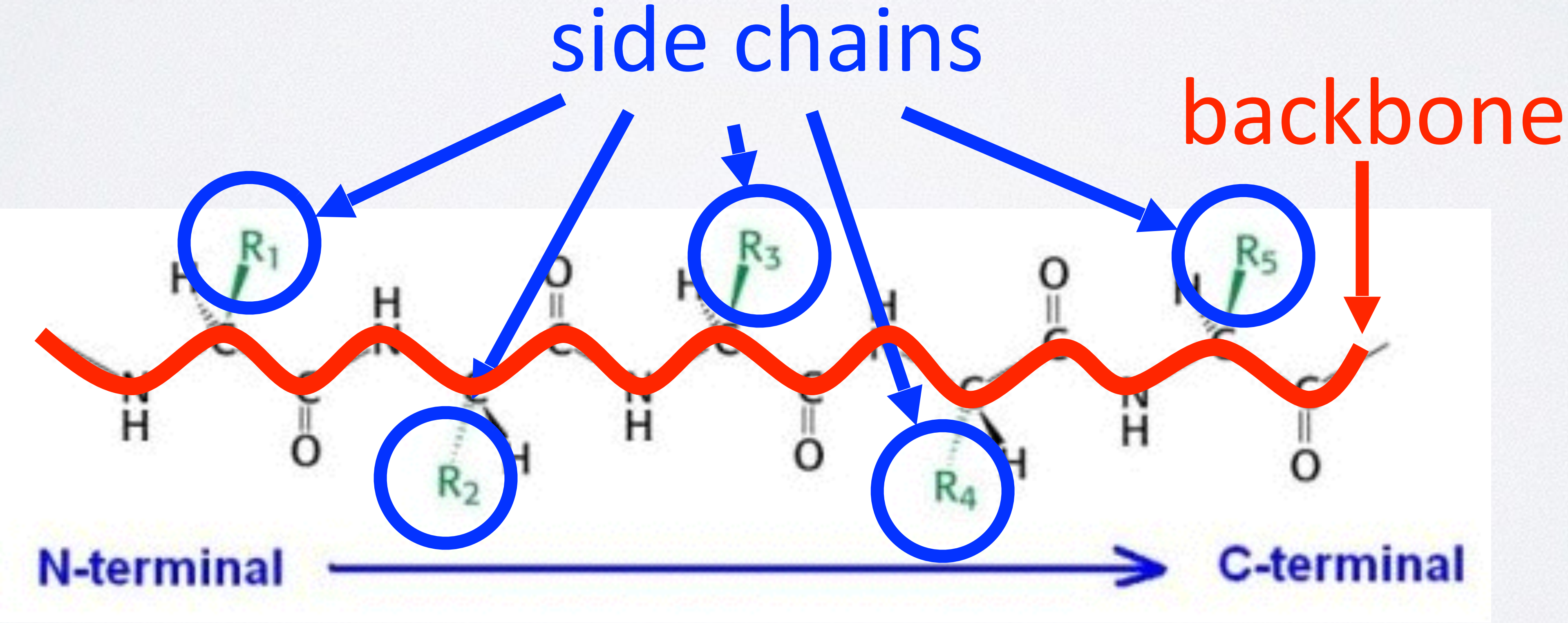
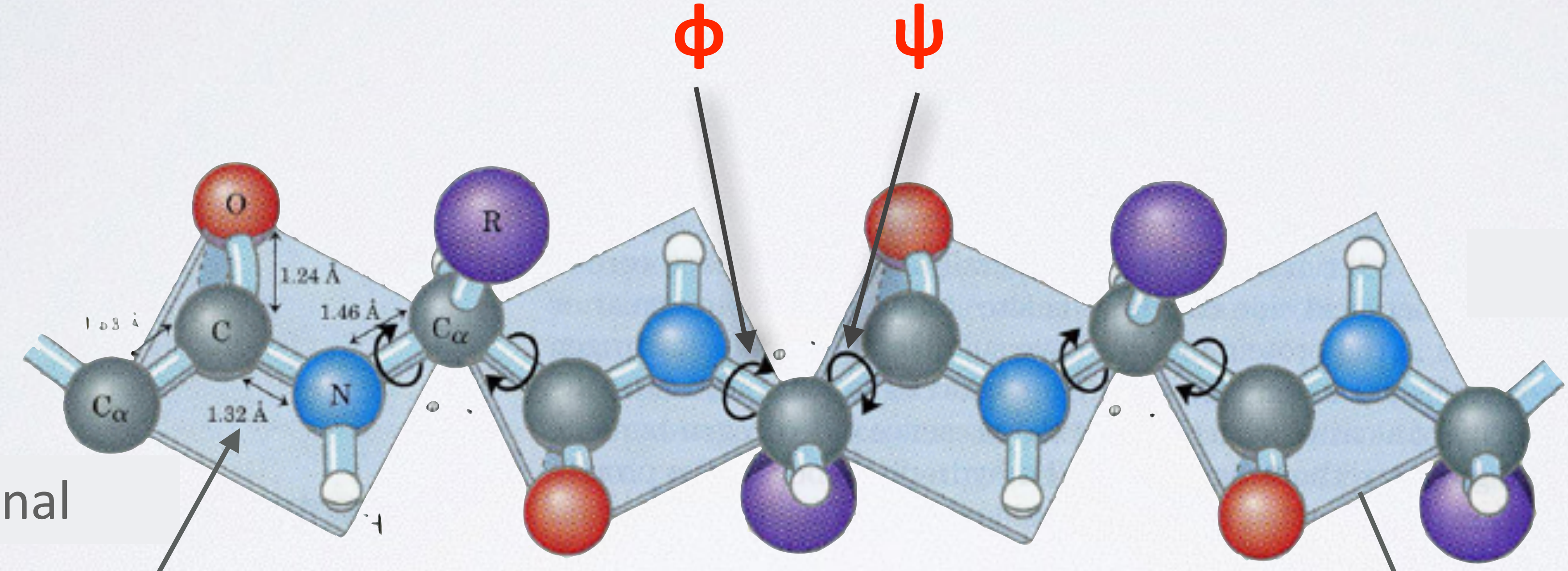


Image from: http://www.ncbi.nlm.nih.gov/books/NBK21581/

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



C-terminal

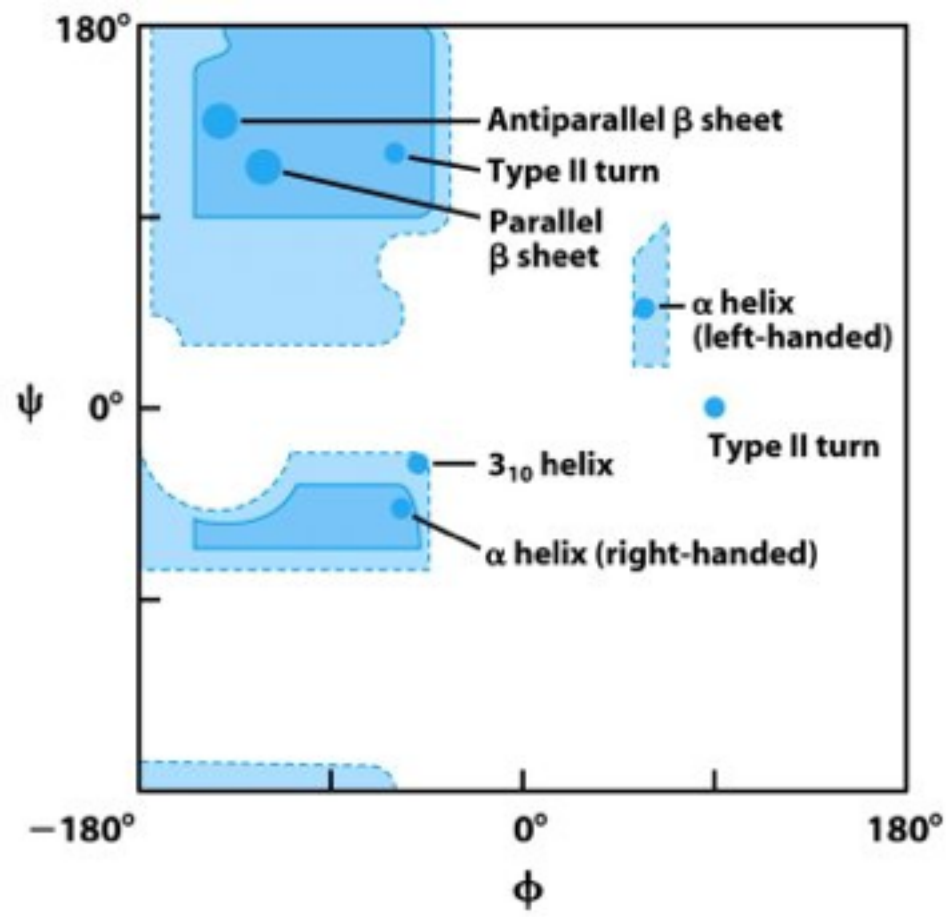
N-terminal

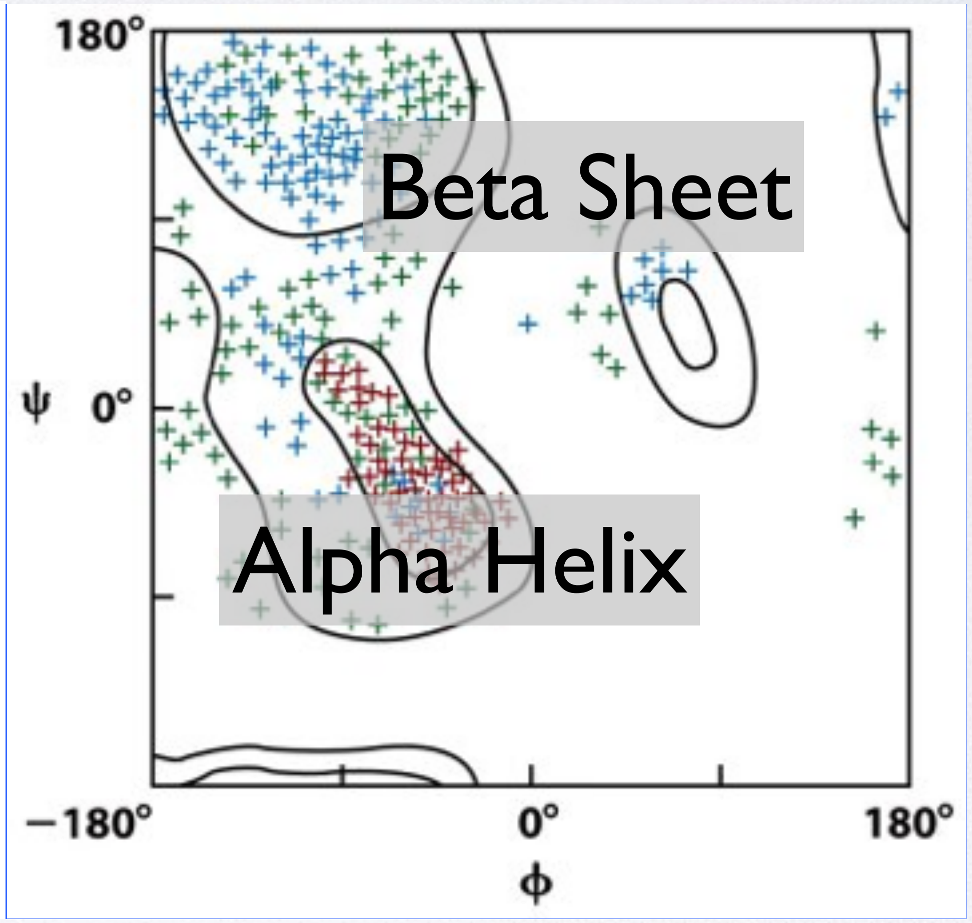
Bond angles and lengths are largely invariant

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**





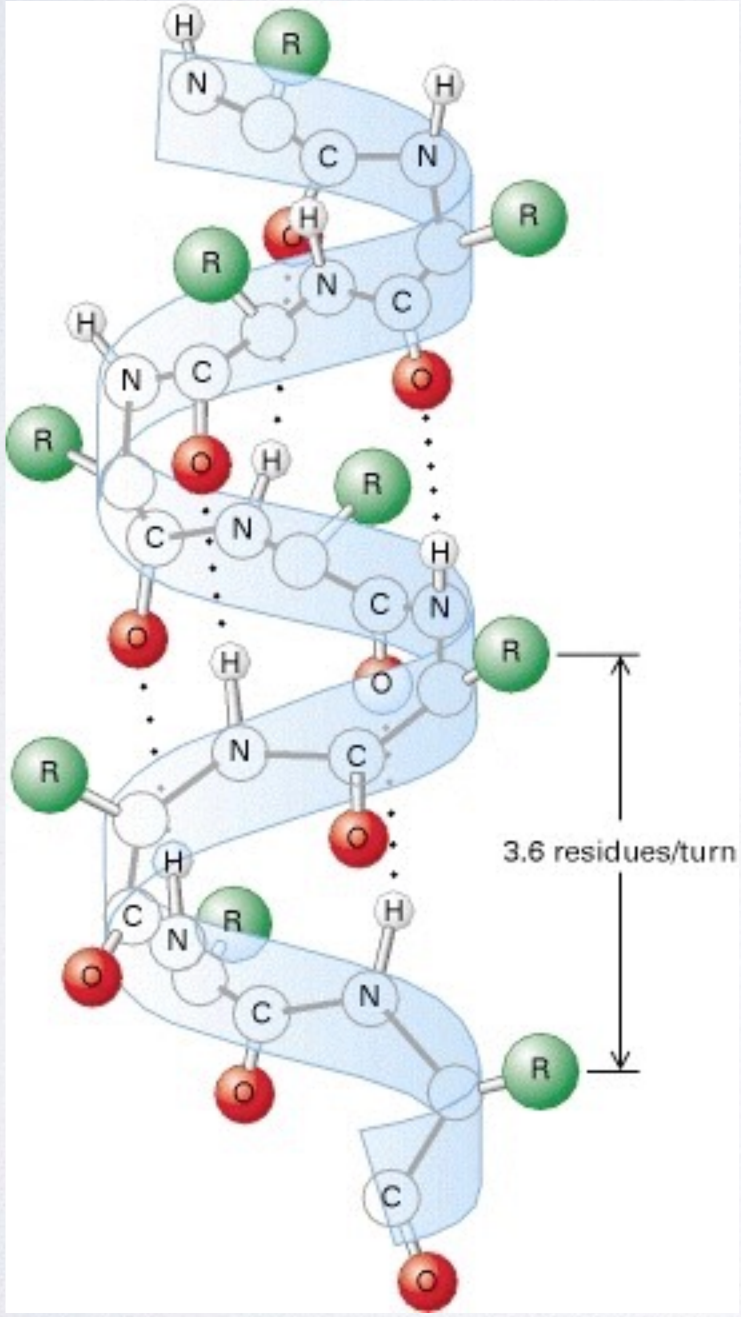
Beta Sheet

Alpha Helix

- 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



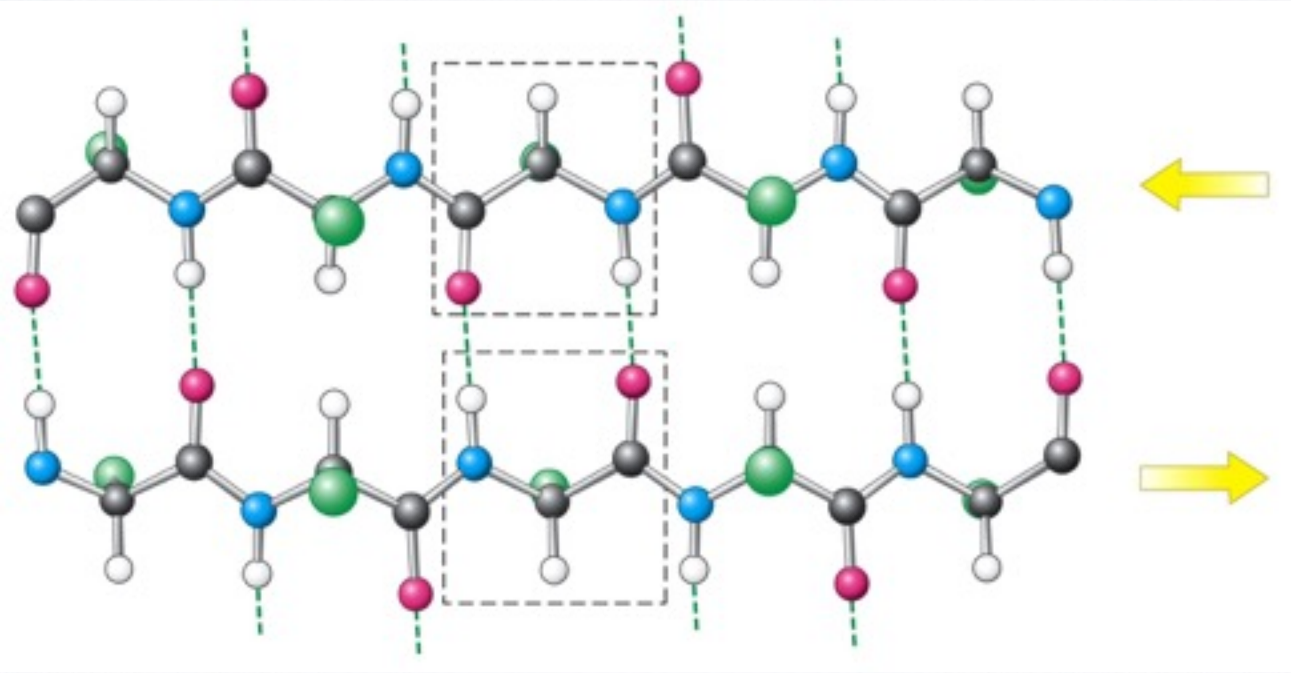
α-helix β-sheets

- 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**

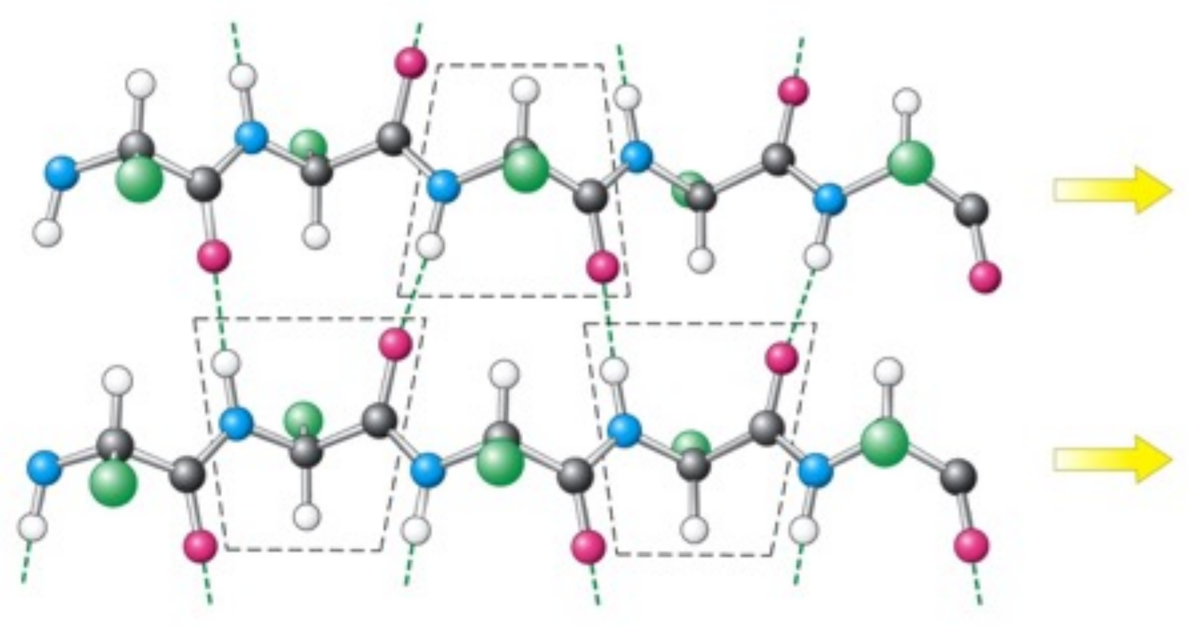


In **antiparallel β-sheets**

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

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

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

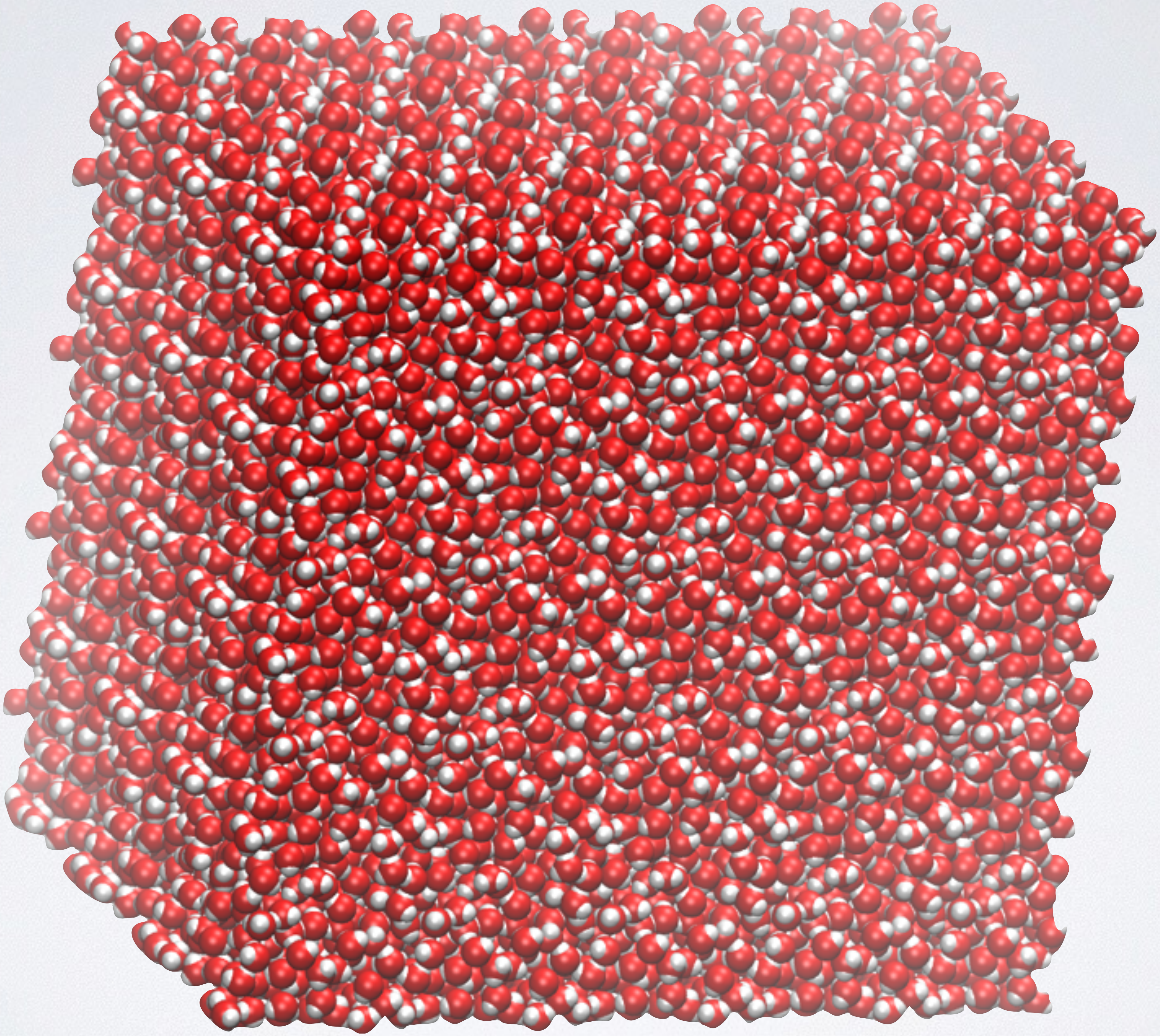


In **parallel β-sheets**

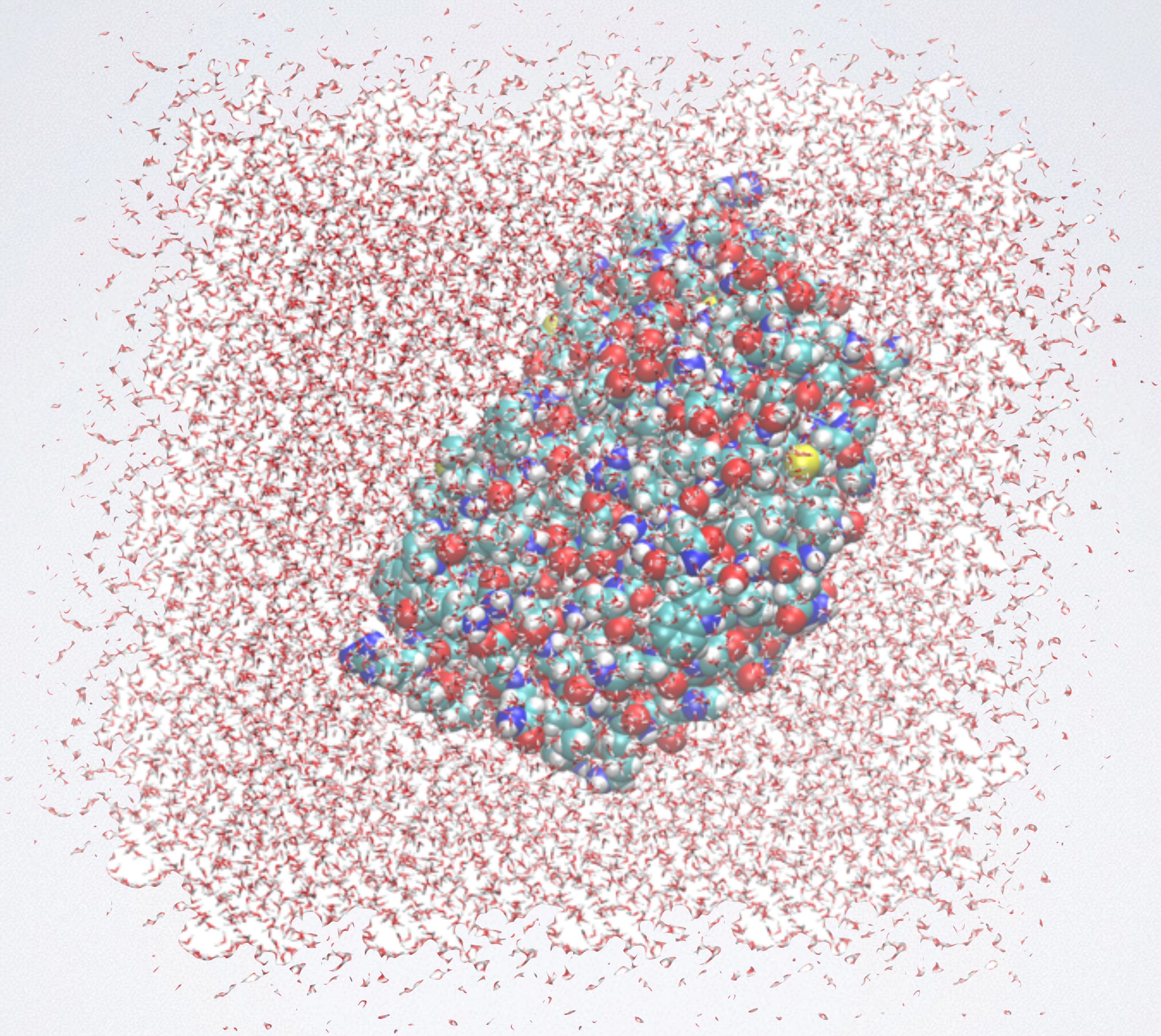
- Adjacent β-strands run in <u>same</u> direction
- Hydrogen bonds (dashed lines) between NH and CO stabilize the structure
- The side chains (in green) are above and below the sheet

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

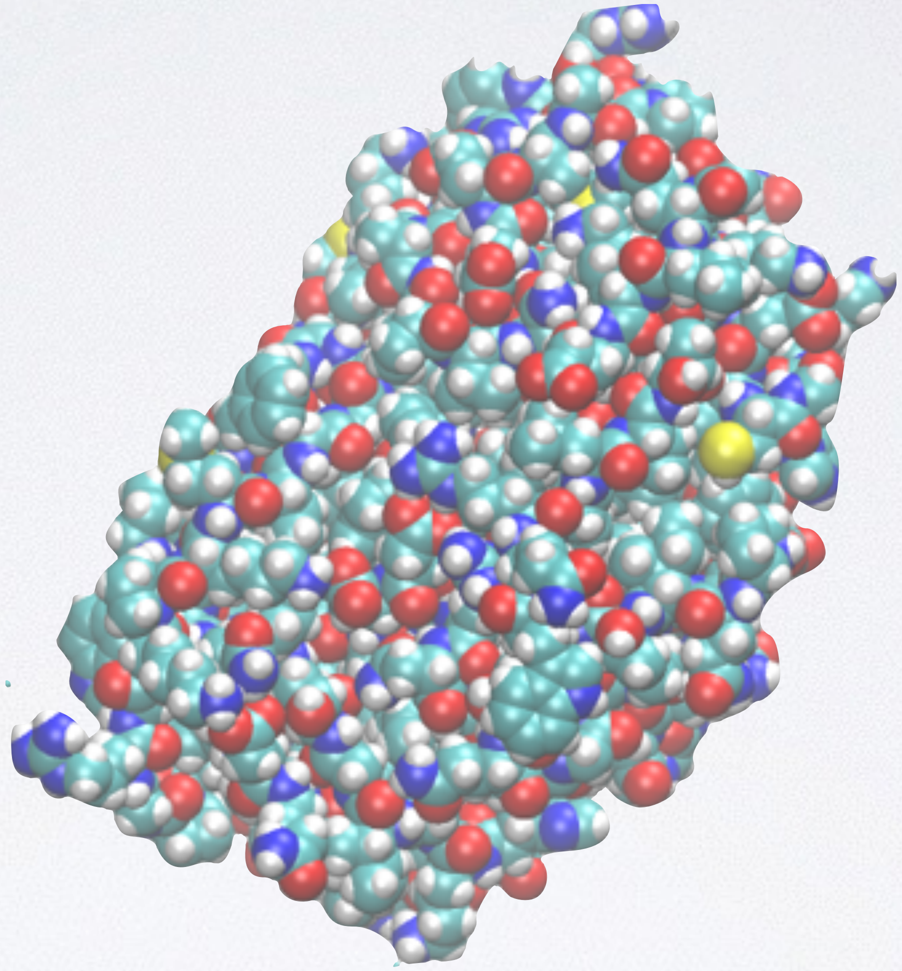
What Does a Protein Look like?



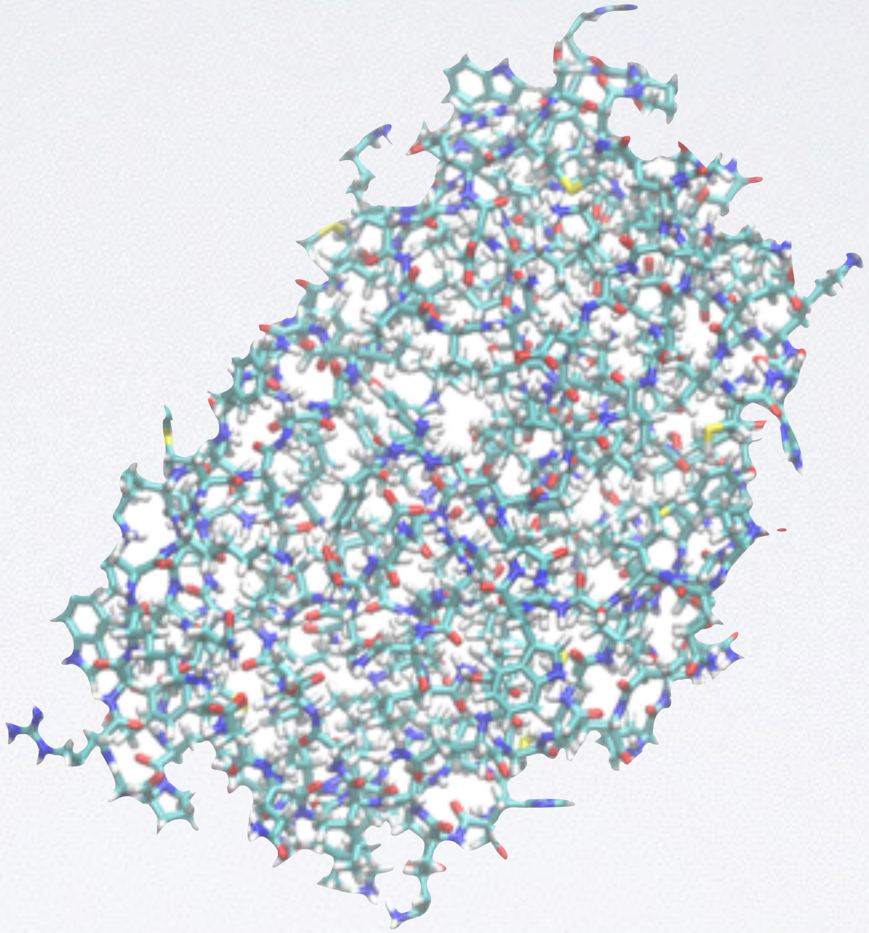
- Proteins are stable (and hidden) in water



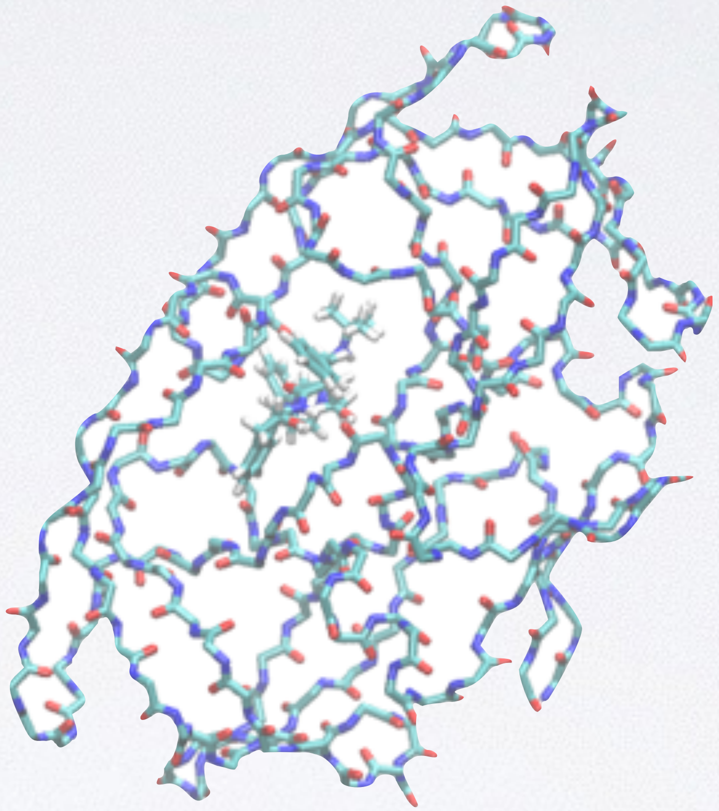
- Proteins closely interact with water



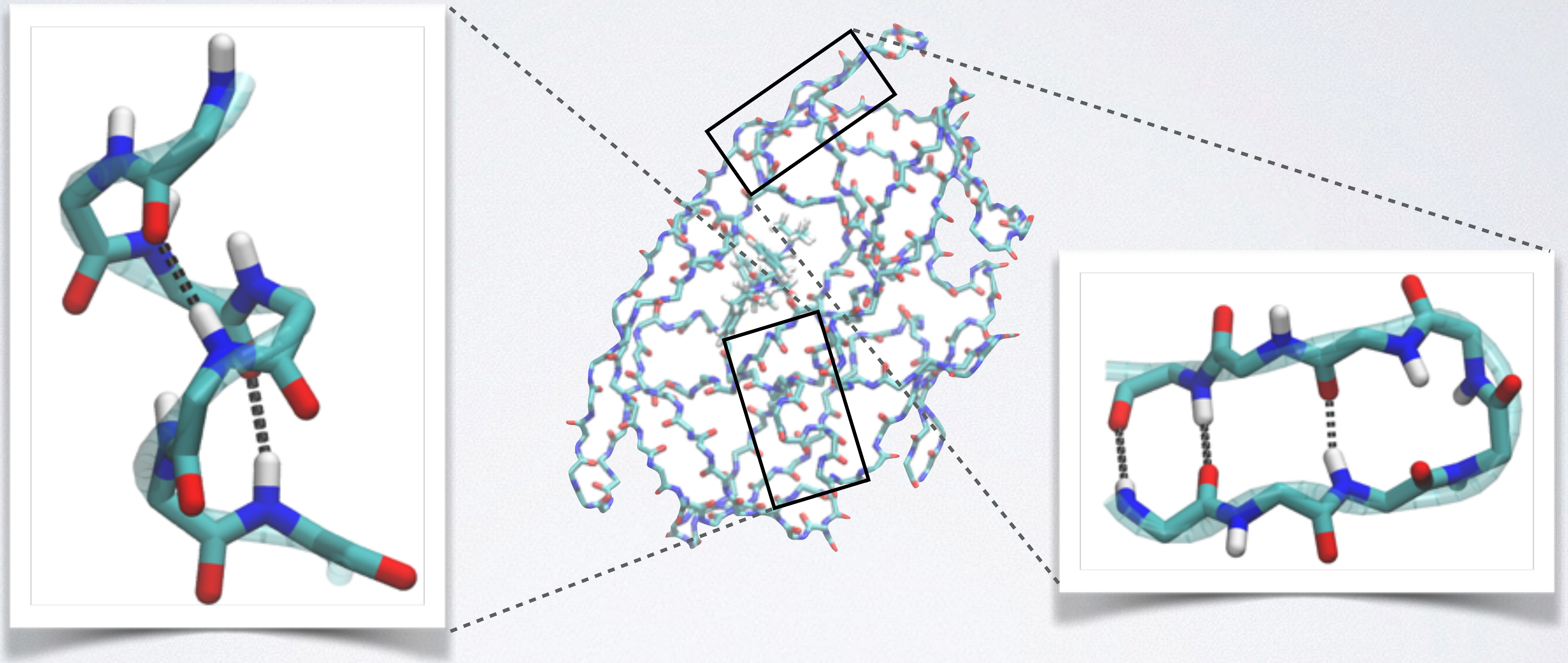
- Proteins are close packed solid but flexible objects (globular)



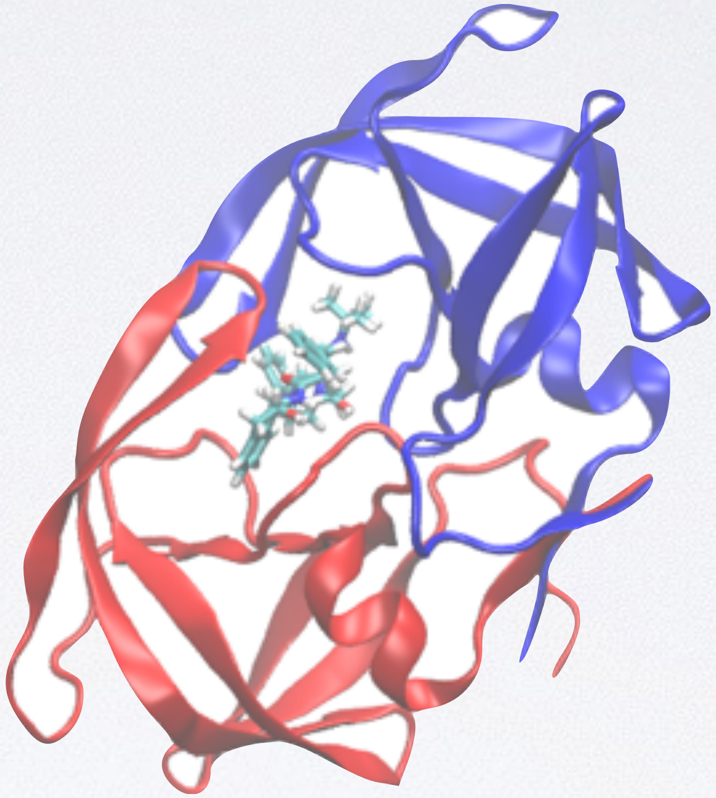
- Due to their large size and complexity it is often hard to see whats important in the structure



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

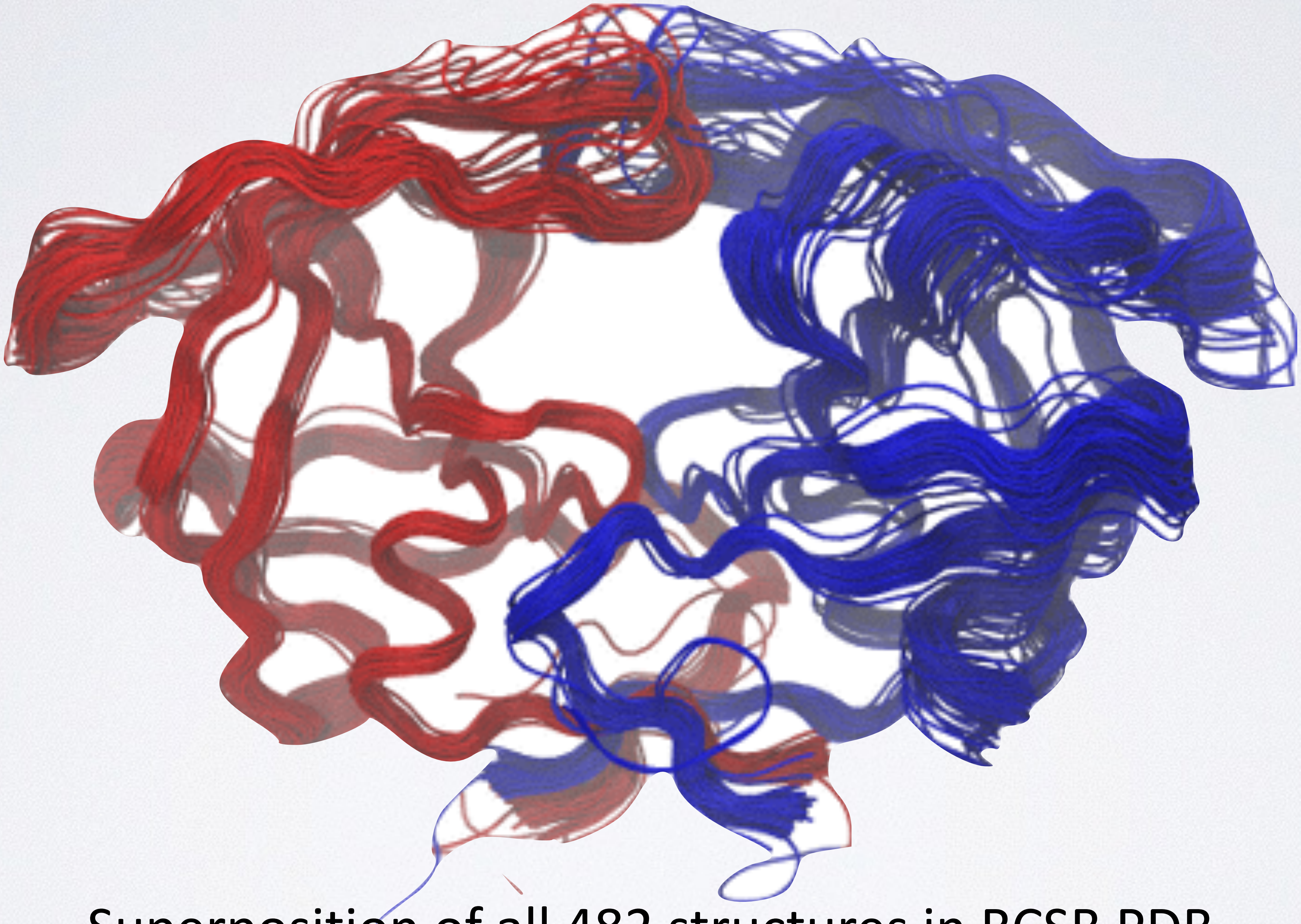


- Backbone or main-chain representation can help trace chain topology & reveal secondary structure



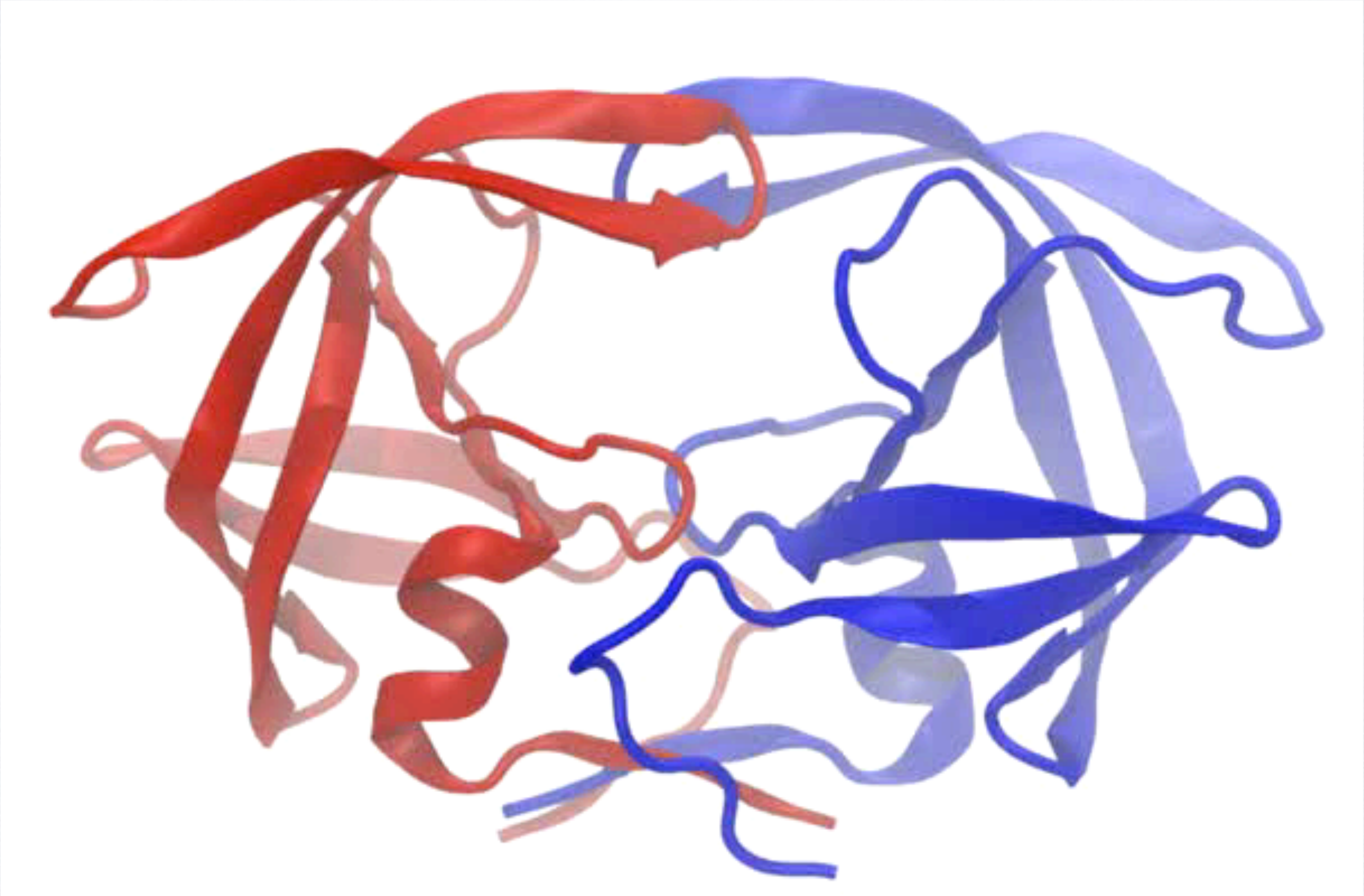
- Simplified secondary structure representations are commonly used to communicate structural details
- Now we can clearly see 2^o, 3^o and 4^o structure
- Coiled chain of connected secondary structures

DISPLACEMENTS REFLECT INTRINSIC FLEXIBILITY



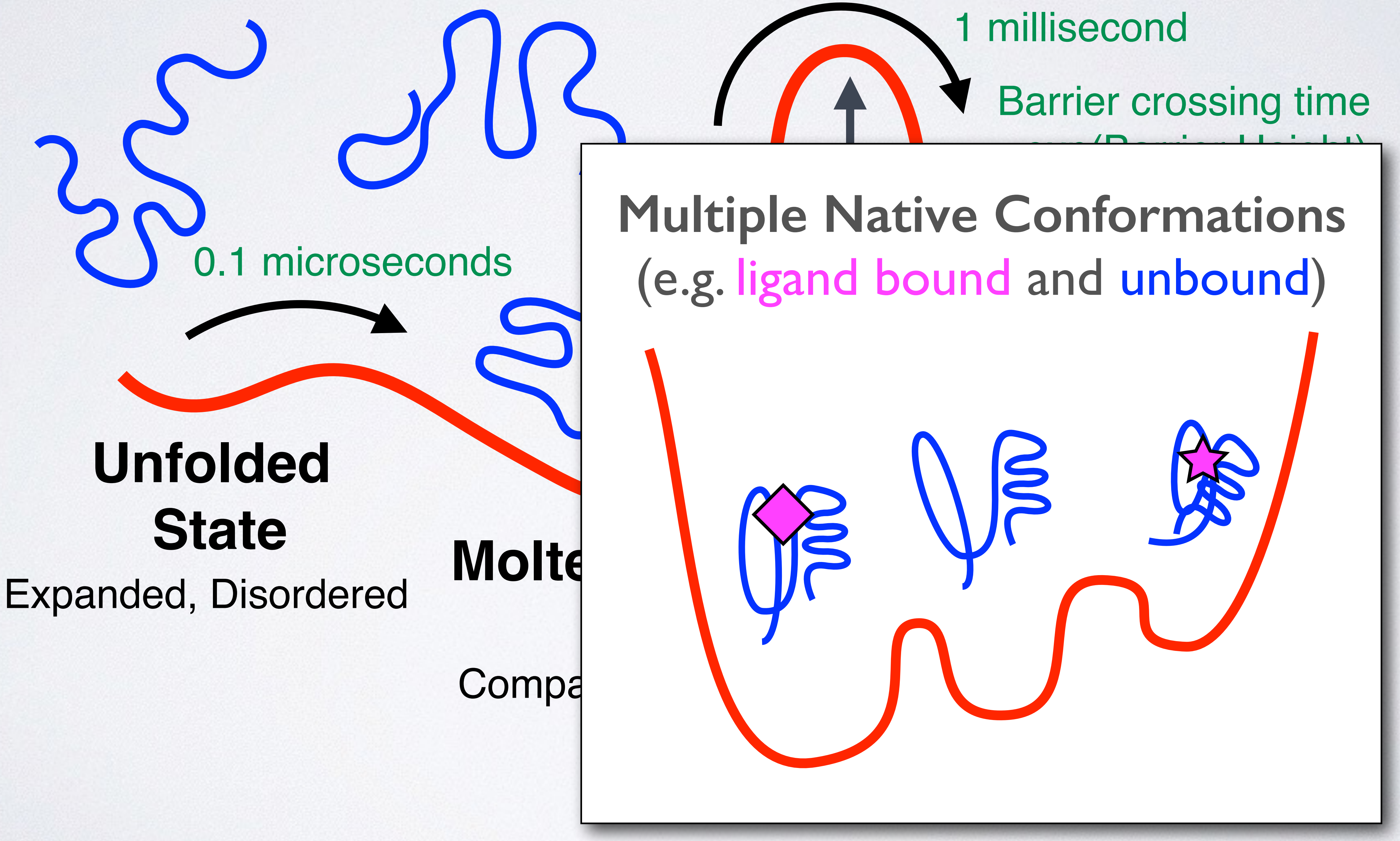
Superposition of all 482 structures in RCSB PDB
(23/09/2015)

DISPLACEMENTS REFLECT INTRINSIC FLEXIBILITY



Principal component analysis (PCA) of experimental structures

KEY CONCEPT: **ENERGY LANDSCAPE**



1 millisecond

Barrier crossing time

exp(Barrier Height)

0.1 microseconds

Unfolded State

Expanded, Disordered

Molte

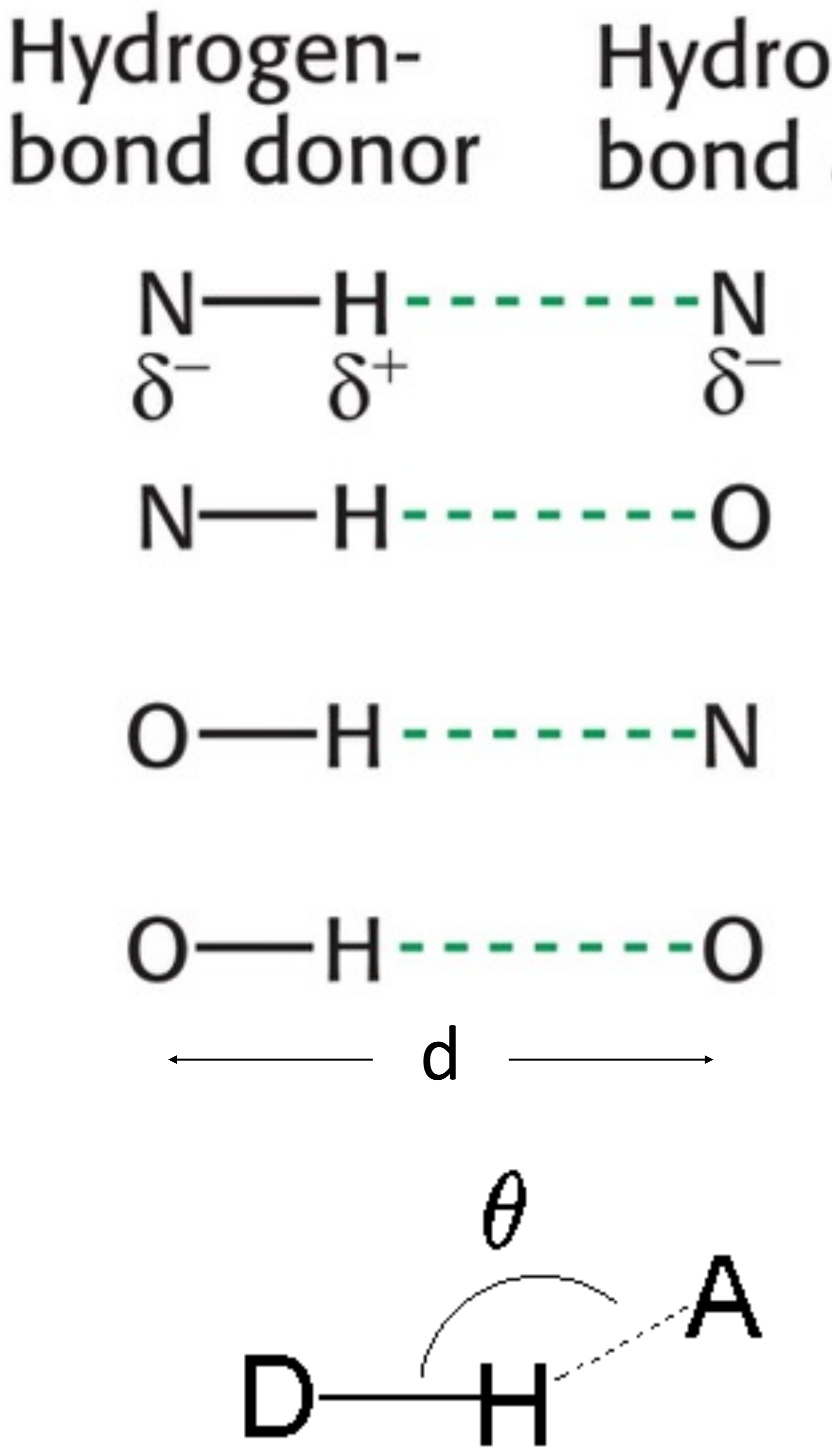
Compa

Multiple Native Conformations

(e.g. ligand bound and unbound)

Key forces affecting structure:

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

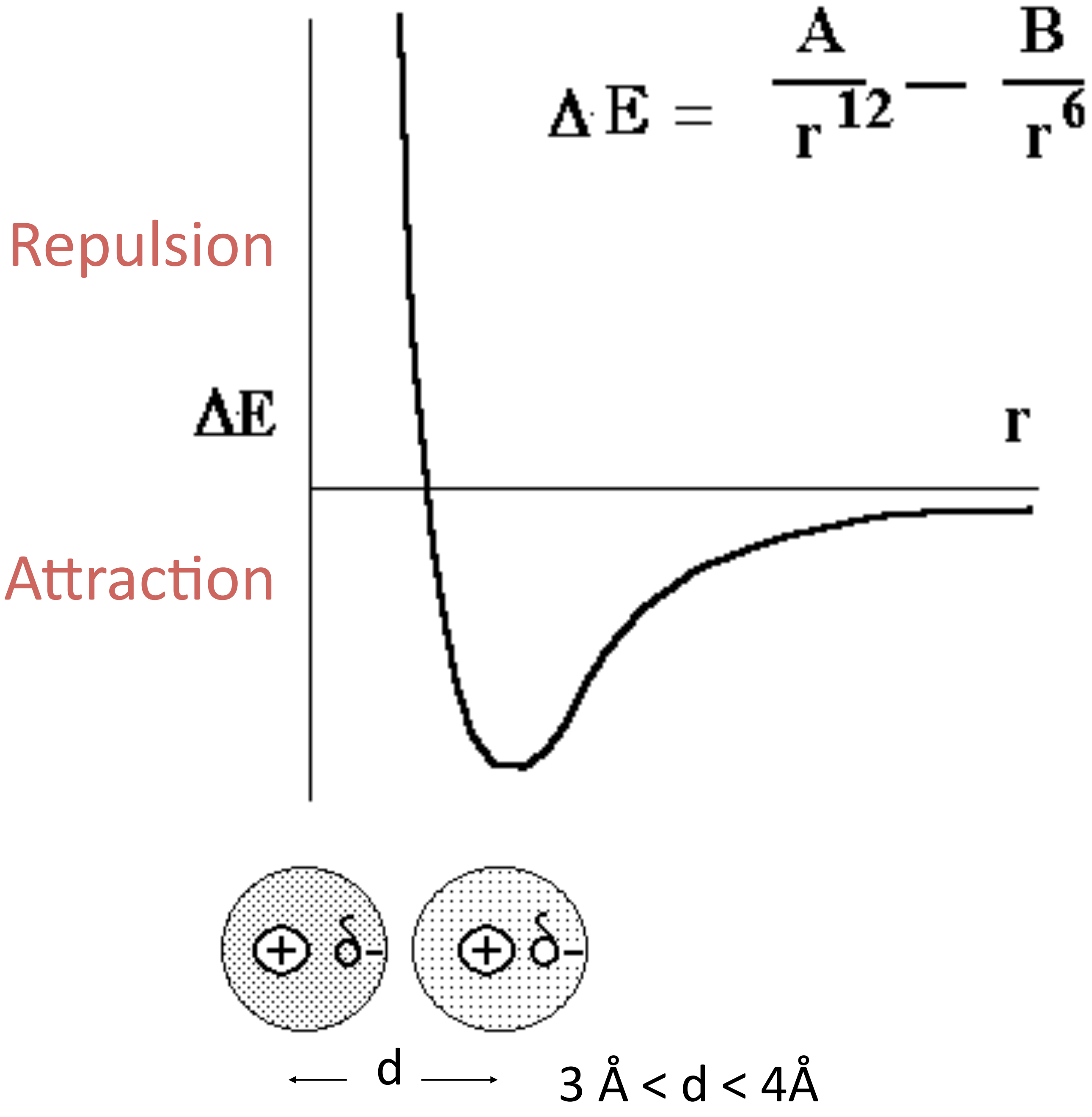


2.6 Å < d < 3.1Å

150° < θ < 180°

Key forces affecting structure:

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



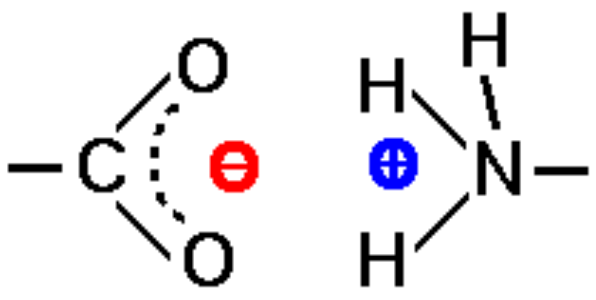
$$\Delta E = \frac{A}{r^{12}} - \frac{B}{r^{6}}$$

3 Å < d < 4Å

Key forces affecting structure:

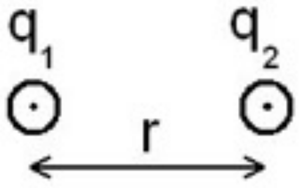
- H-bonding
- Van der Waals
- Electrostatics
- Hydrophobicity
- Disulfide Bridges

← d → 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

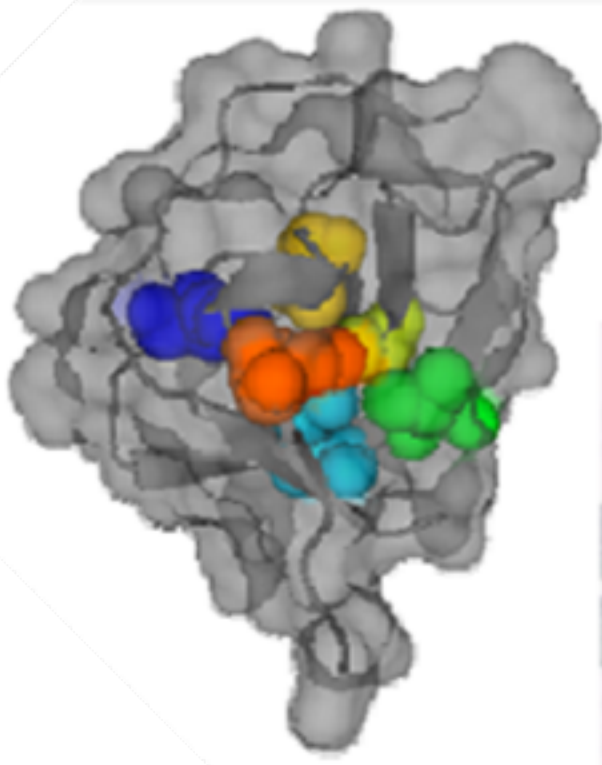
D = Dielectric constant (vacuum = 1; H_2O = 80)

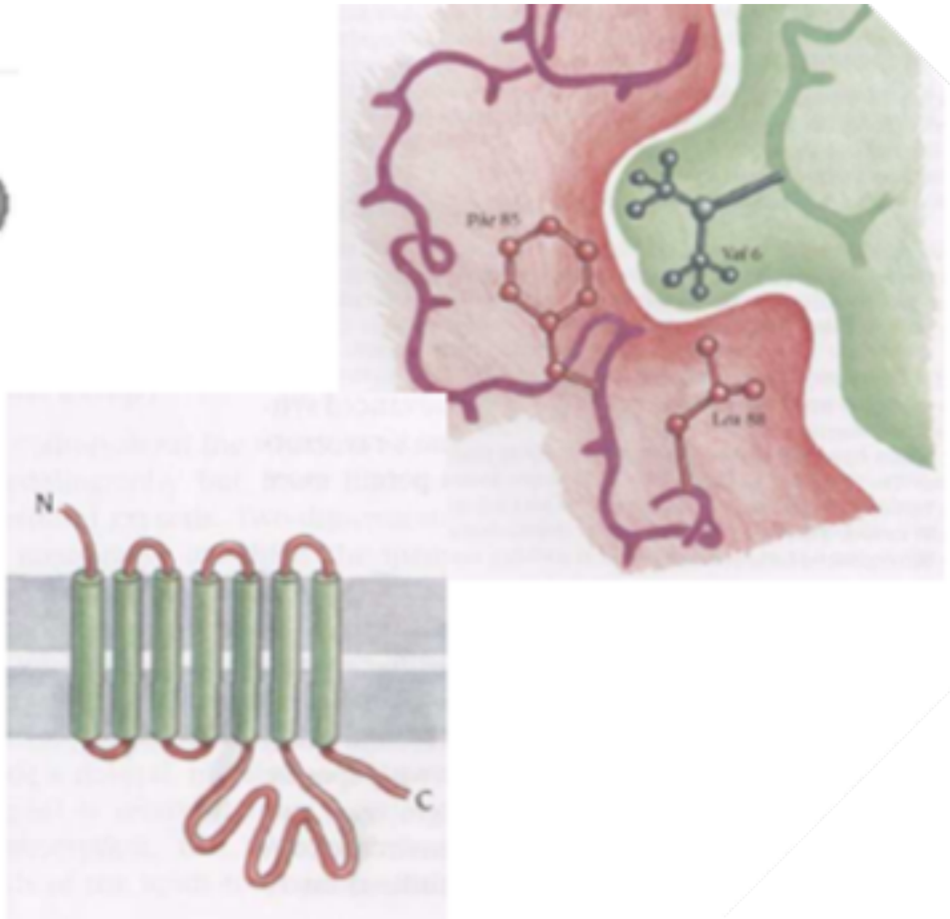
q_1 & q_2 = electronic charges (Coulombs)

r = distance (Å)

Key forces affecting structure:

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



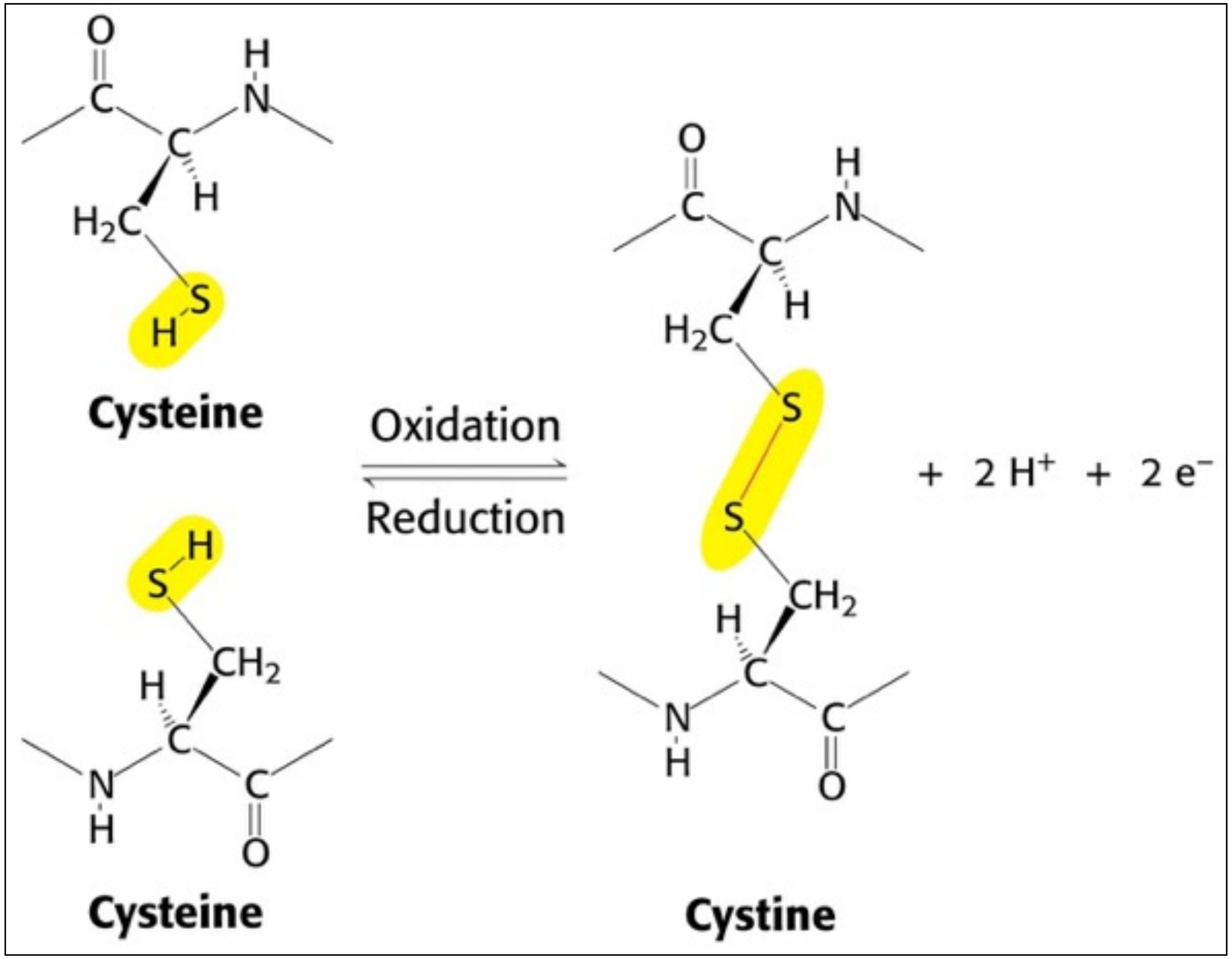


The force that causes hydrophobic molecules or nonpolar portions of molecules to aggregate together rather than to dissolve in water is called Hydrophobicity (*Greek, "water fearing"*). This is not a separate bonding force; rather, it is the result of the energy required to insert a nonpolar molecule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

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

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:

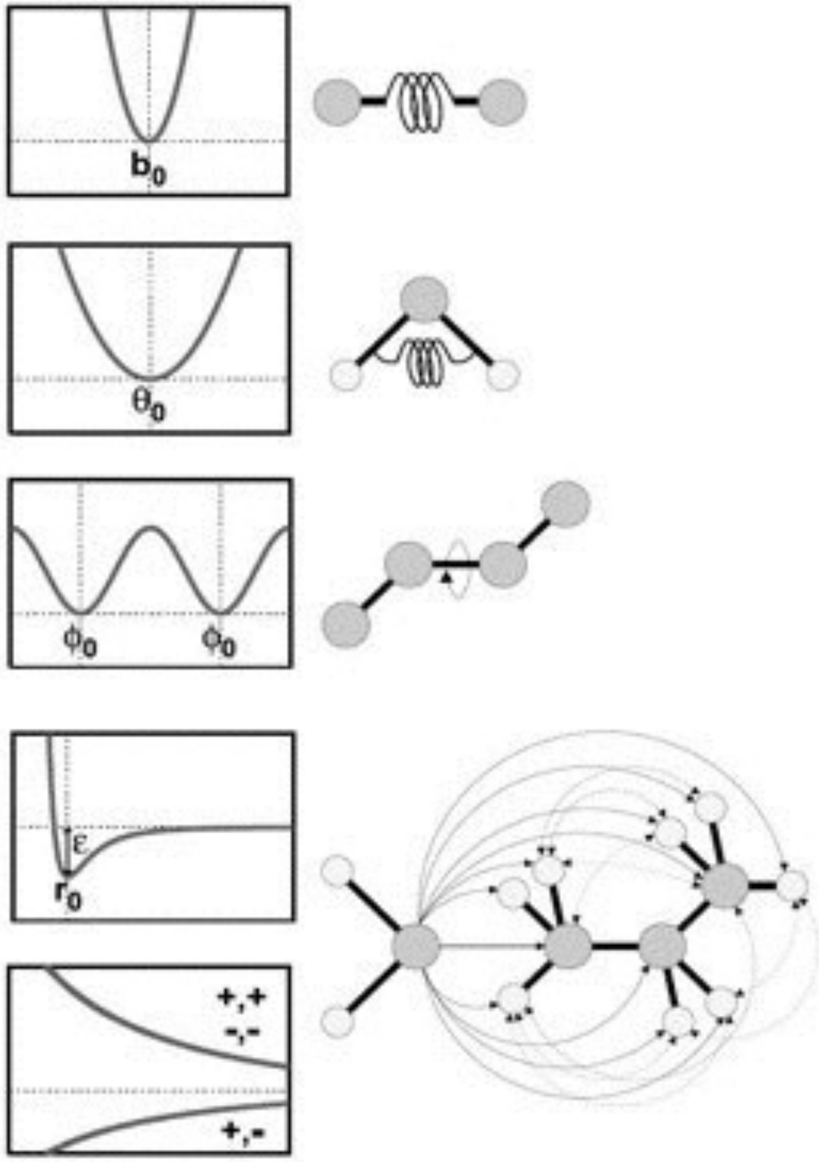
(1). **Physics-Based**

(2). **Knowledge-Based**

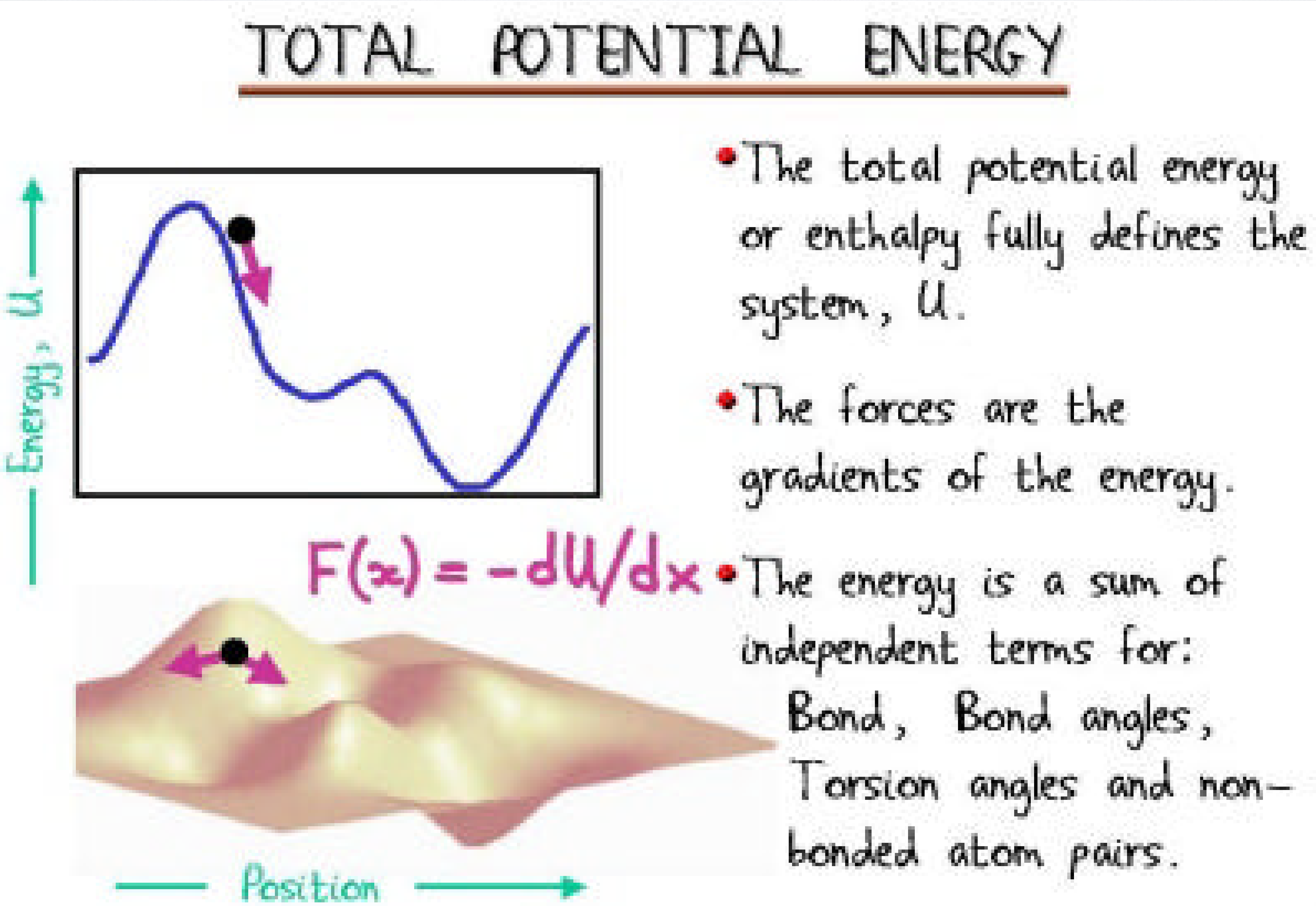
PHYSICS-BASED POTENTIALS
ENERGY TERMS FROM PHYSICAL THEORY

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

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/

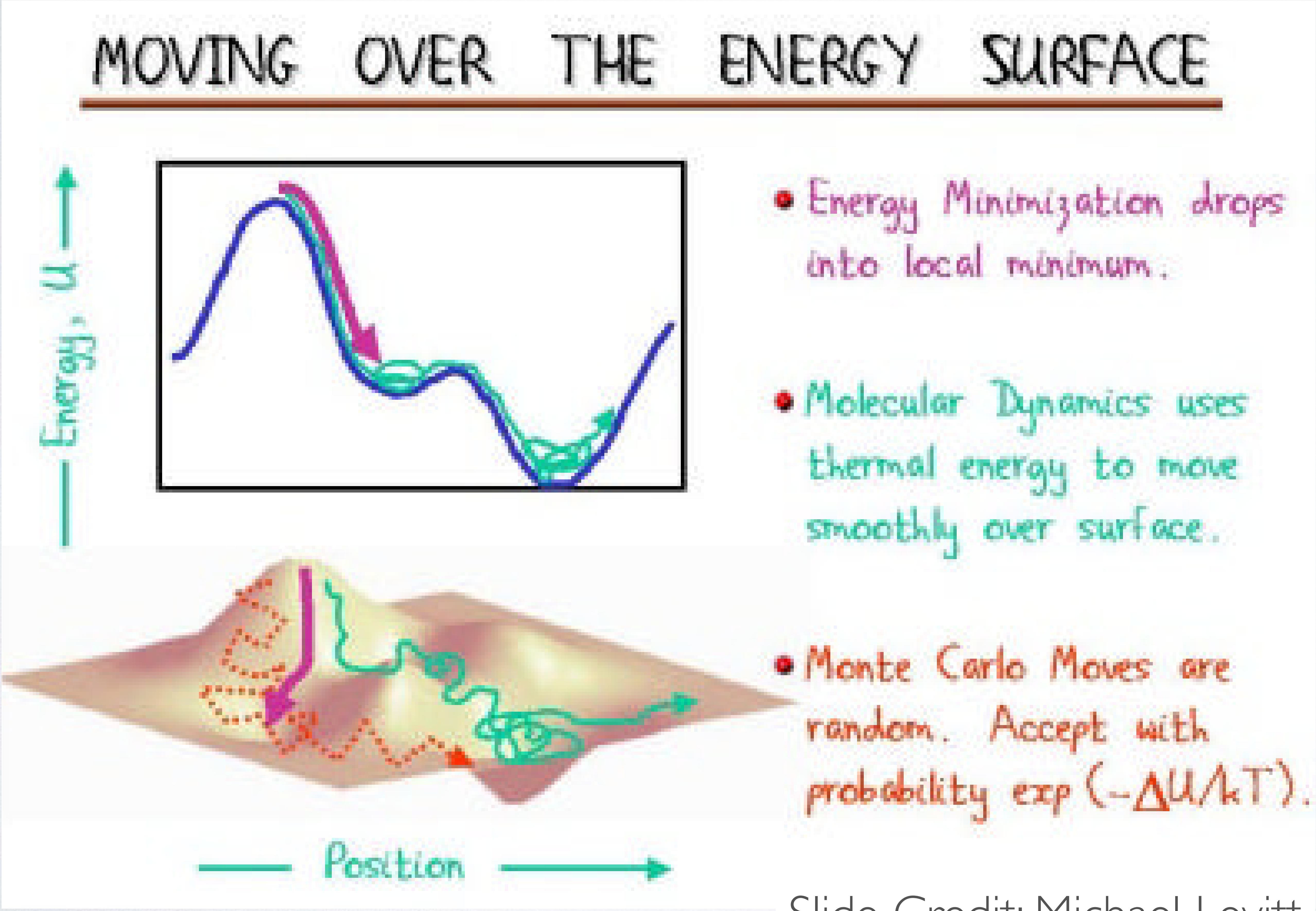


TOTAL POTENTIAL ENERGY

$$F(x) = -dU/dx$$

- 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.

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(−ΔU/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

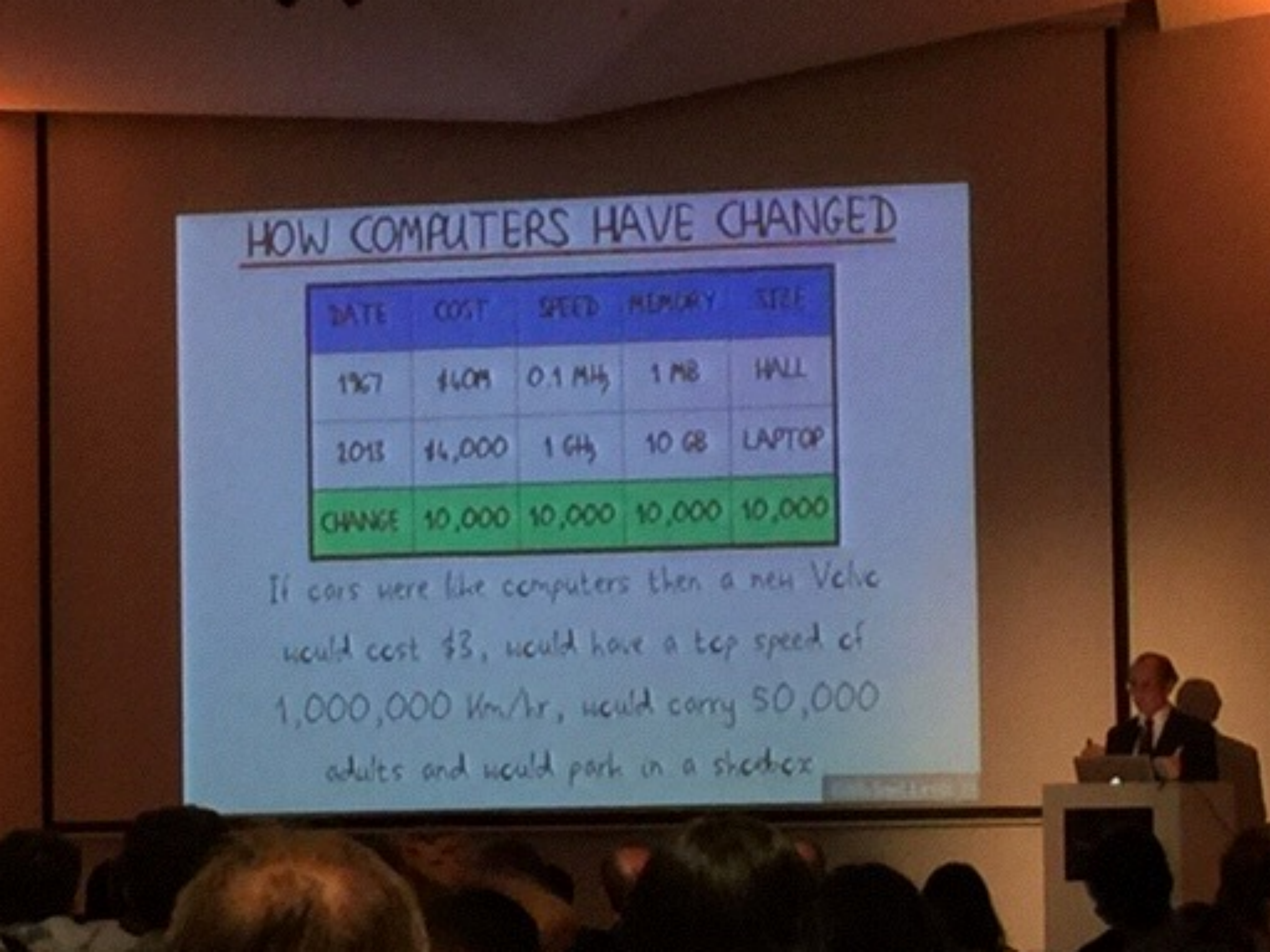
Status

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

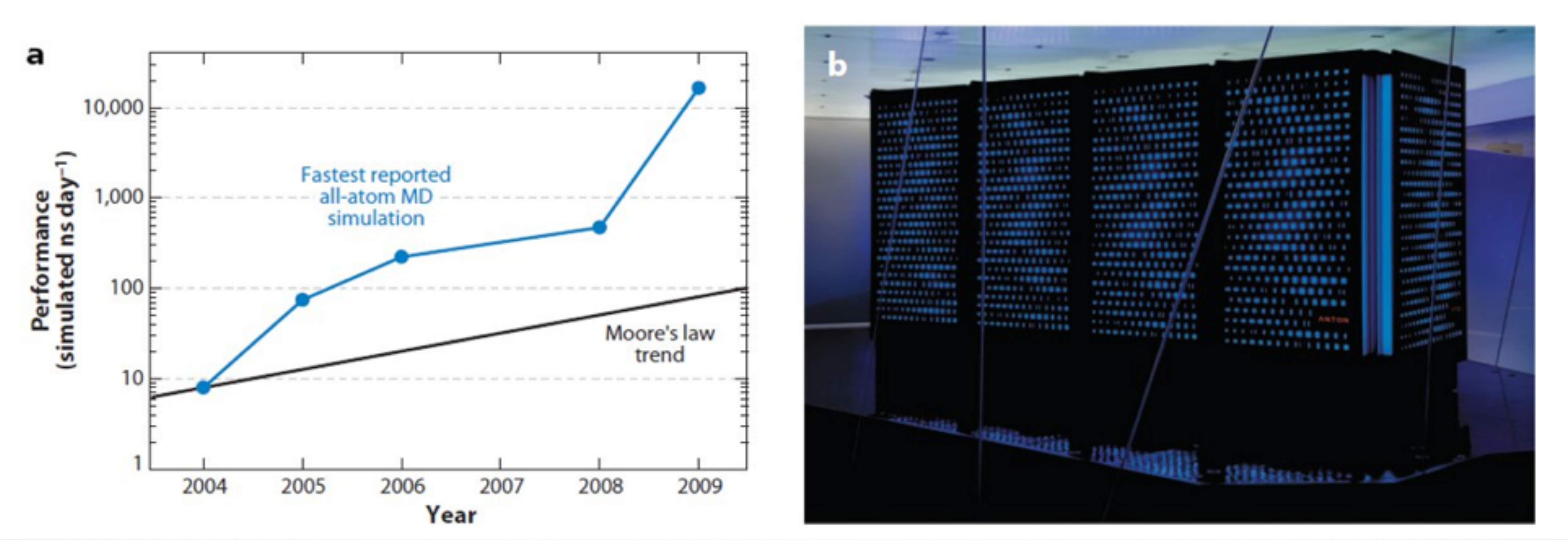


HOW COMPUTERS HAVE CHANGED

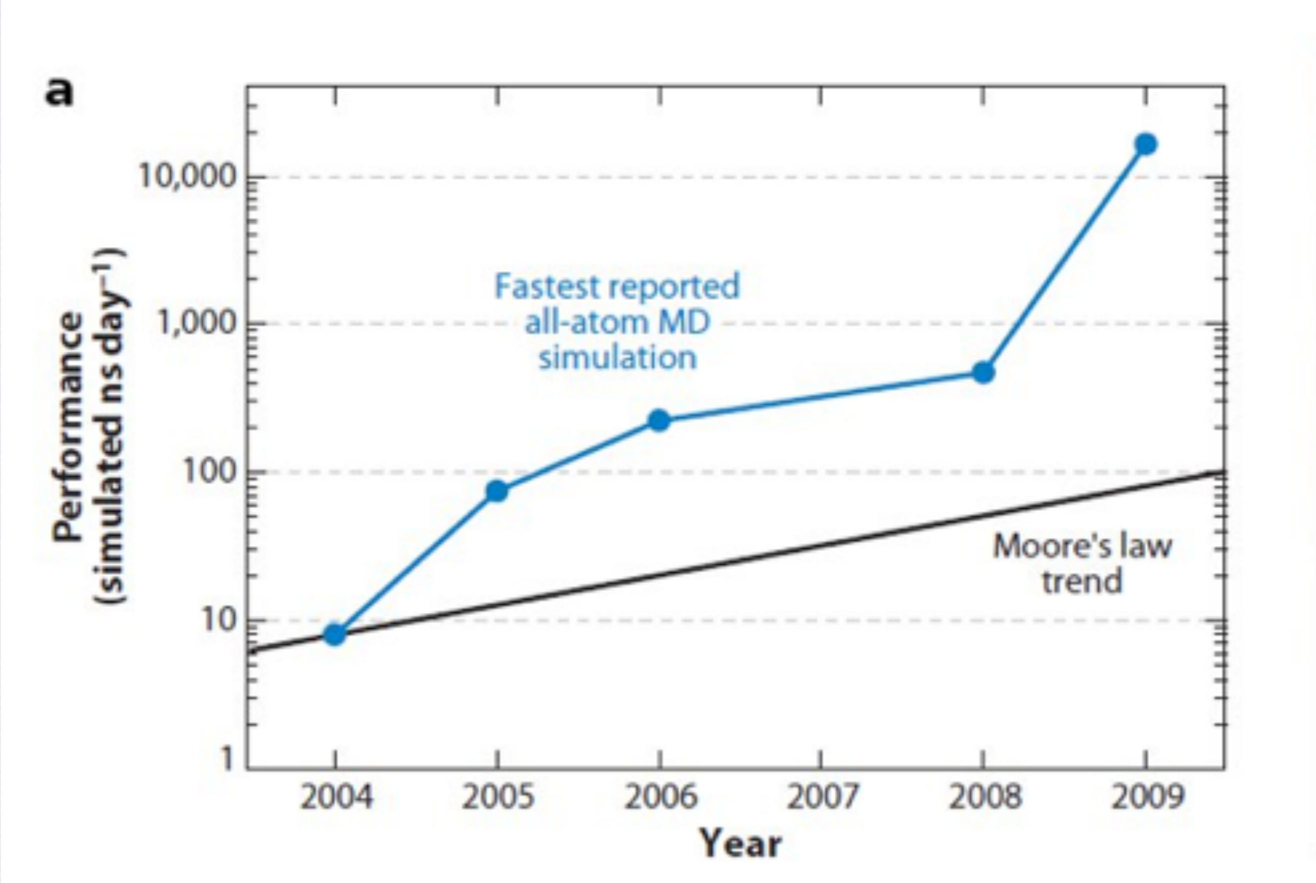
DATE	COST	SPEED	MEMORY	SIZE
1967	$40M	0.1 MHz	1 MB	HALL
2013	$4,000	1 GHz	10 GB	LAPTOP
CHANGE	10,000	10,000	10,000	10,000

If cars were like 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



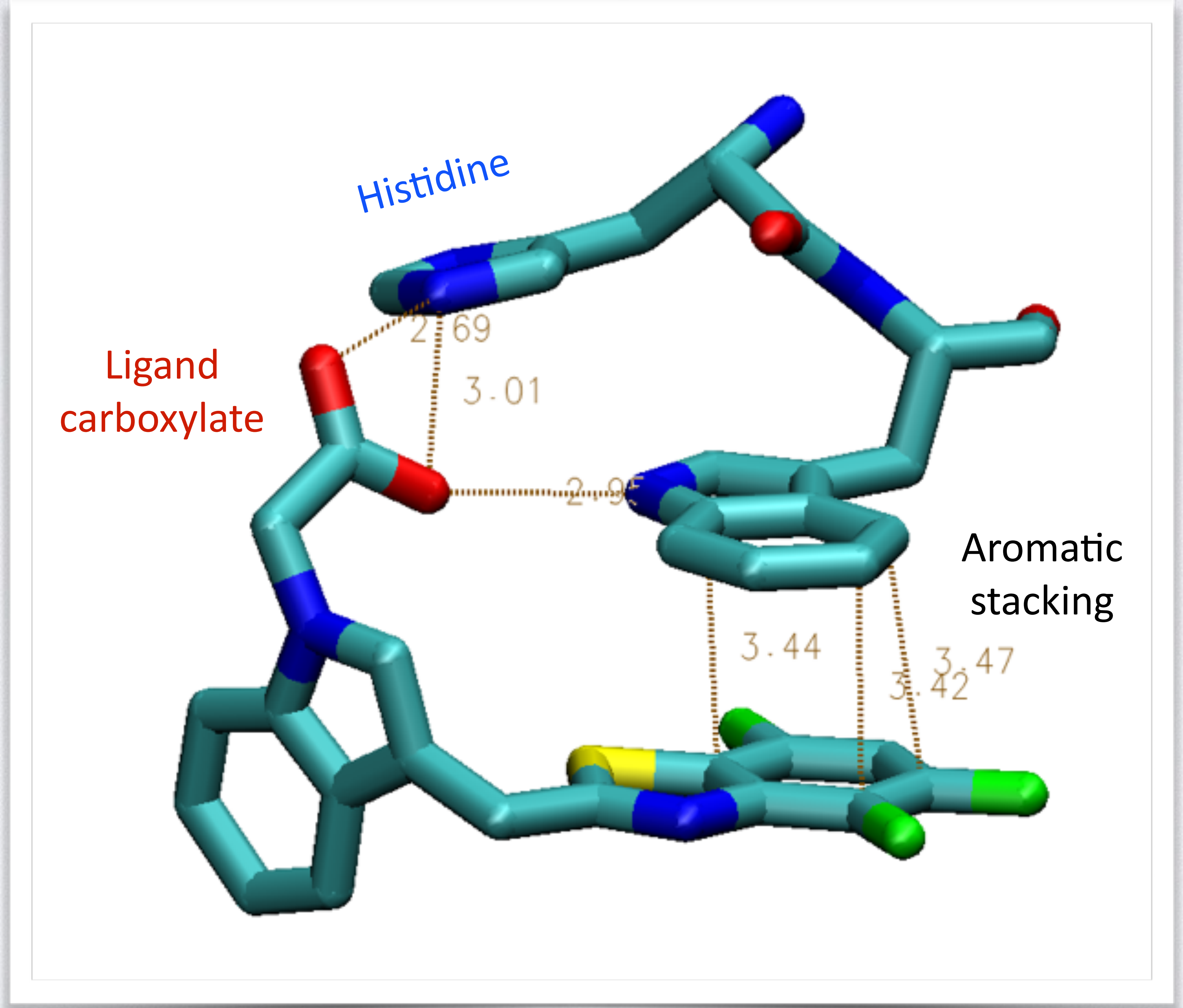
KEY CONCEPT: POTENTIAL FUNCTIONS DESCRIBE A SYSTEMS **ENERGY** AS A FUNCTION OF ITS **STRUCTURE**

Two main approaches:

(1). **Physics-Based**

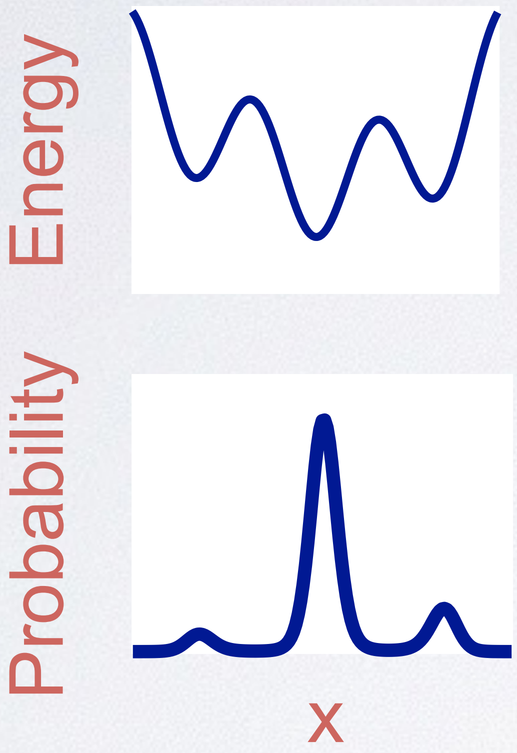
(2). **Knowledge-Based**

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

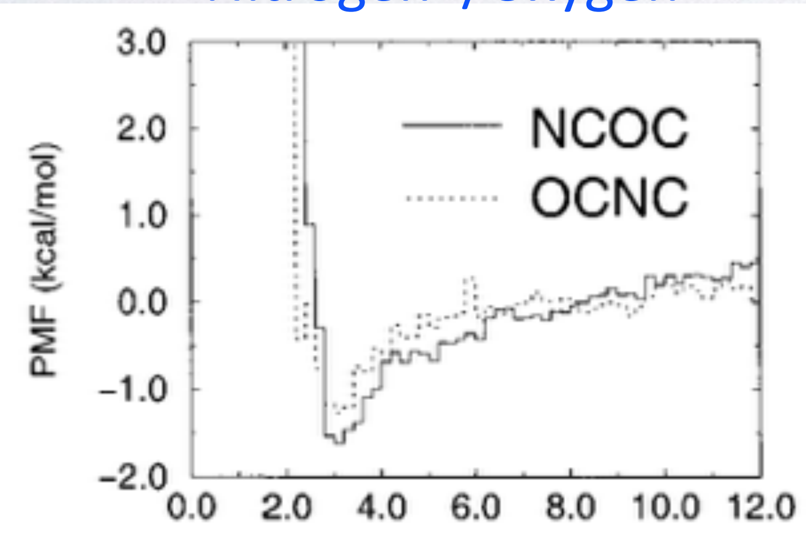
1. For each structure, histogram the distances from O to every histidine N
2. Sum the histograms over all structures to obtain $p(r_{O\text{-}N})$
3. Compute $E(r_{O\text{-}N})$ from $p(r_{O\text{-}N})$

KNOWLEDGE-BASED DOCKING POTENTIALS

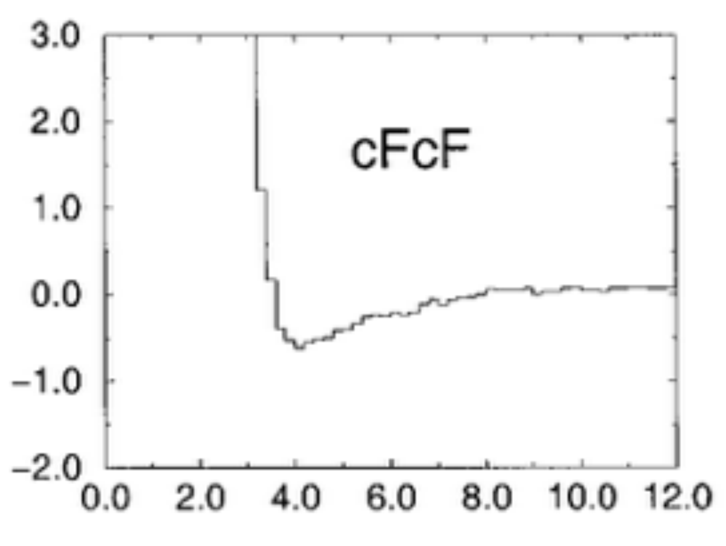
“PMF”, Muegge & Martin, J. Med. Chem. (1999) 42:791

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

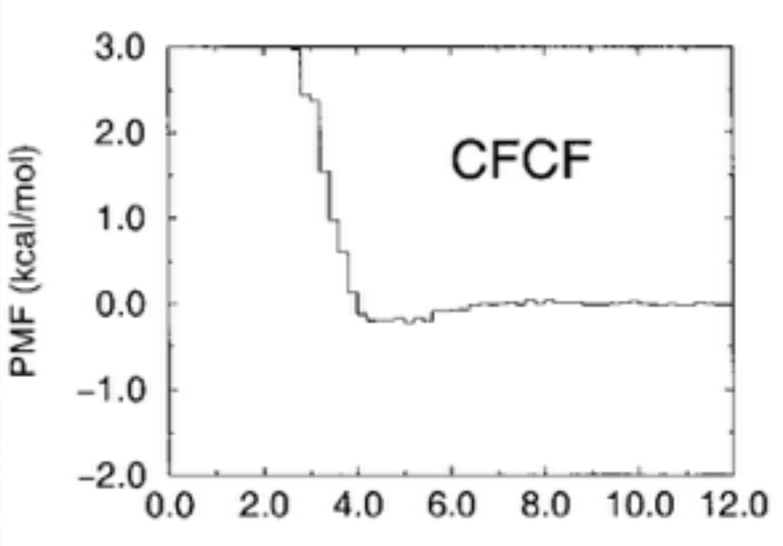
Nitrogen$^+$/Oxygen$^-$



Aromatic carbons



Aliphatic carbons



Atom-atom distance (Angstroms)

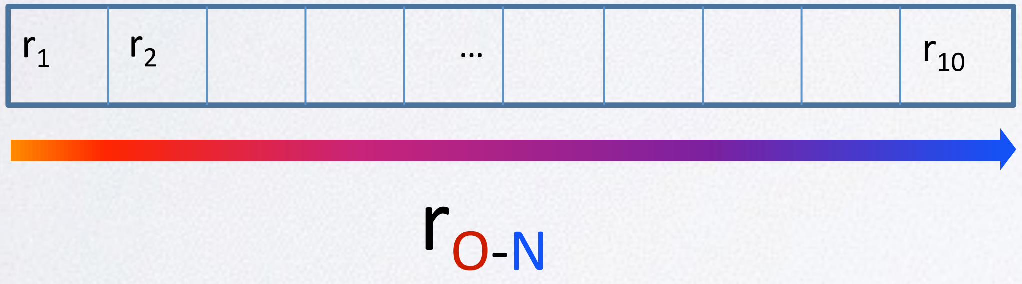
$$E_{prot-lig} = E_{vdw} + \sum_{pairs\,(ij)} E_{type(ij)}(r_{ij})$$

LIMITATIONS OF KNOWLEDGE-BASED POTENTIALS

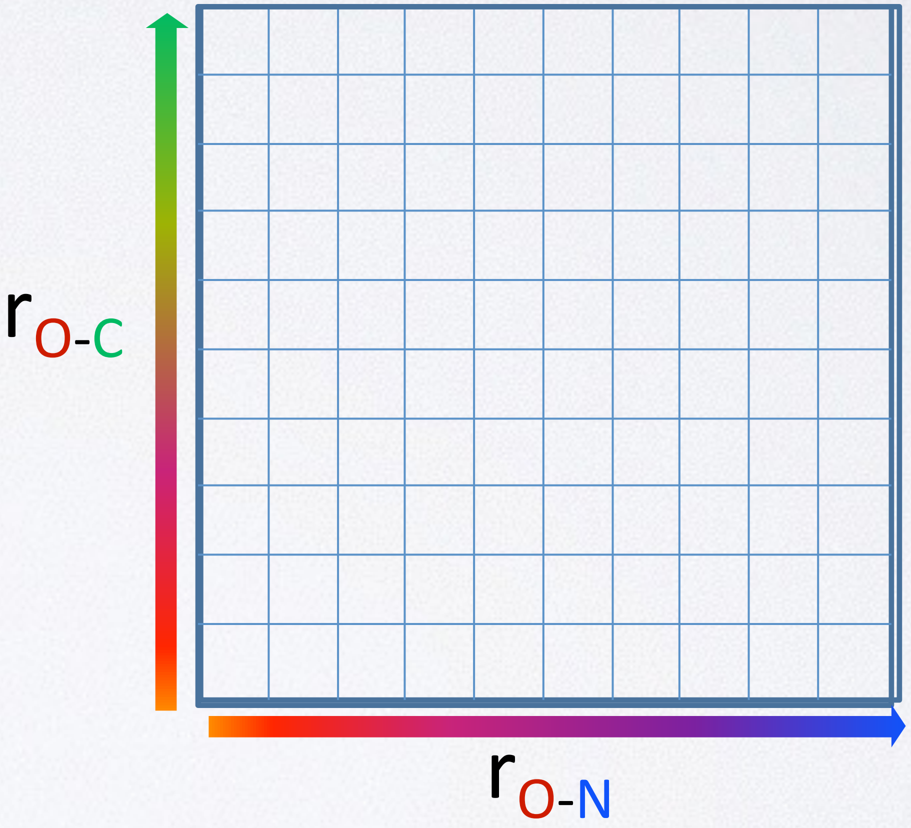
1. Statistical limitations

(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)

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**

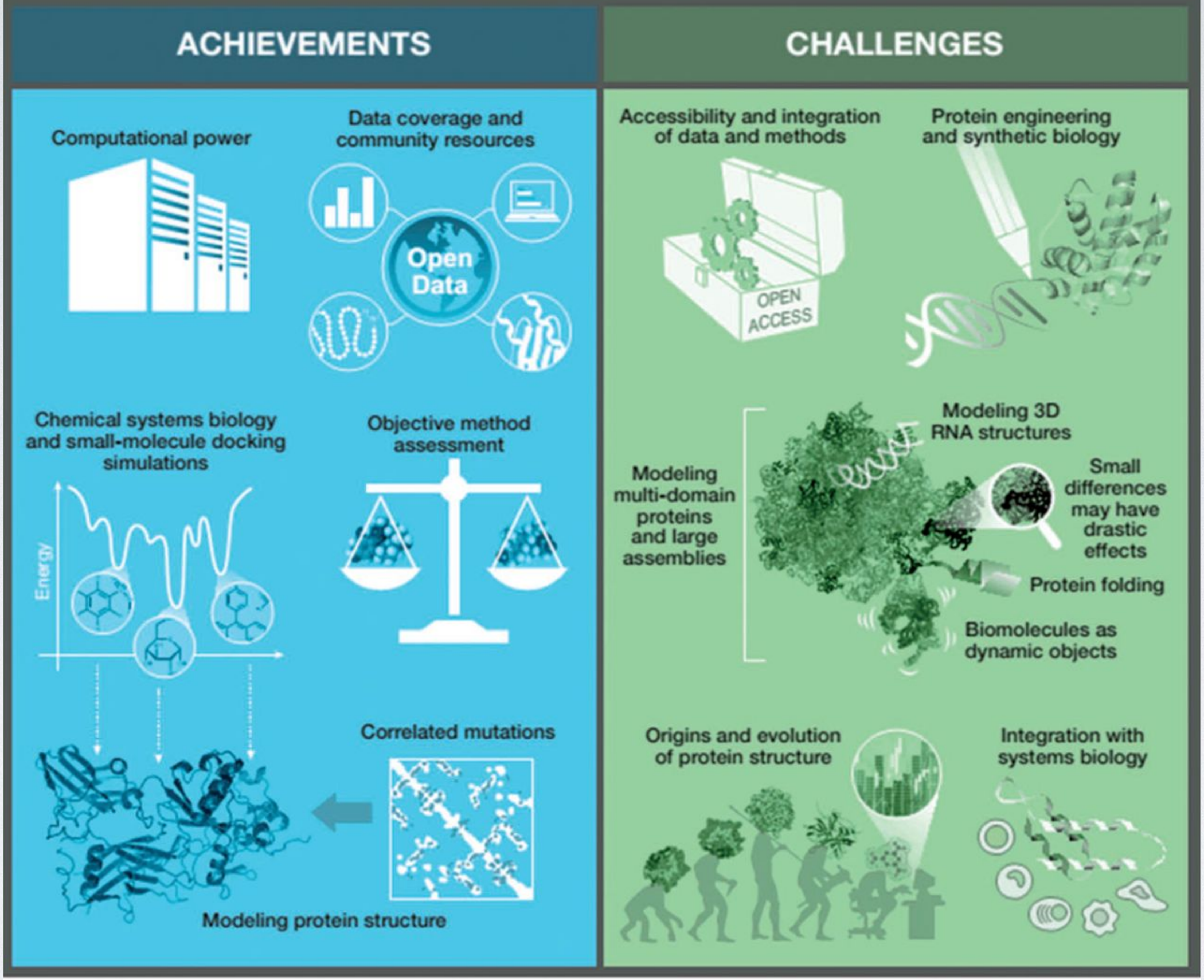
  Questions and hypotheses should always be designed such that they do not depend crucially on the precise numbers used for the various parameters.

- **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.

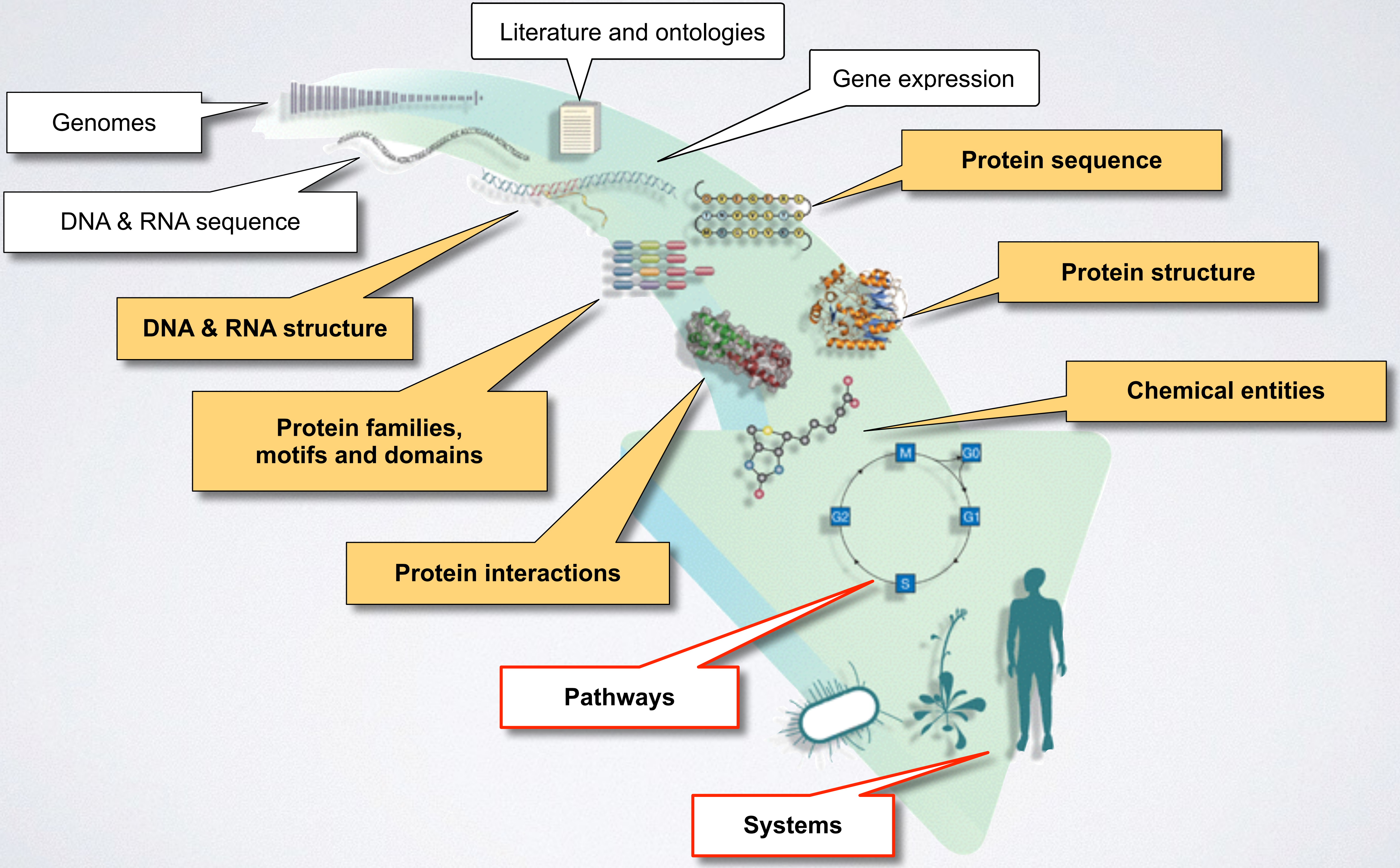
SUMMARY

- 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
Accessibility and integration of data and methods
OPEN ACCESS
Protein engineering and synthetic biology
Chemical systems biology and small-molecule docking simulations
Energy
Objective method assessment
Modeling 3D RNA structures
Modeling multi-domain proteins and large assemblies
Small differences may have drastic effects
Protein folding
Biomolecules as dynamic objects
Correlated mutations
Modeling protein structure
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

Protein sequence

DNA & RNA sequence

Protein structure

DNA & RNA structure

Chemical entities

Protein families, motifs and domains

Protein interactions

Pathways

Systems