BGGN 213

Structural Bioinformatics II

Barry Grant

UC San Diego

http://thegrantlab.org/bggn213

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

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

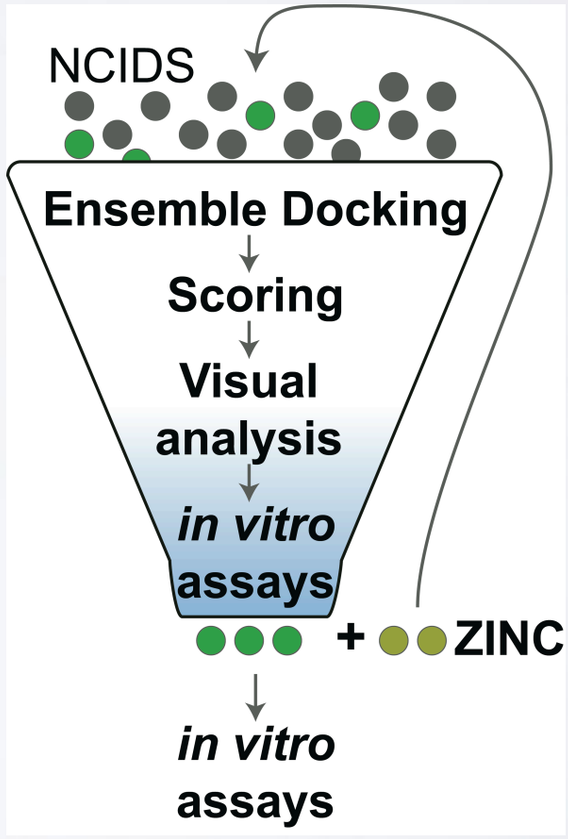
COMPUTER-AIDED LIGAND DESIGN

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**

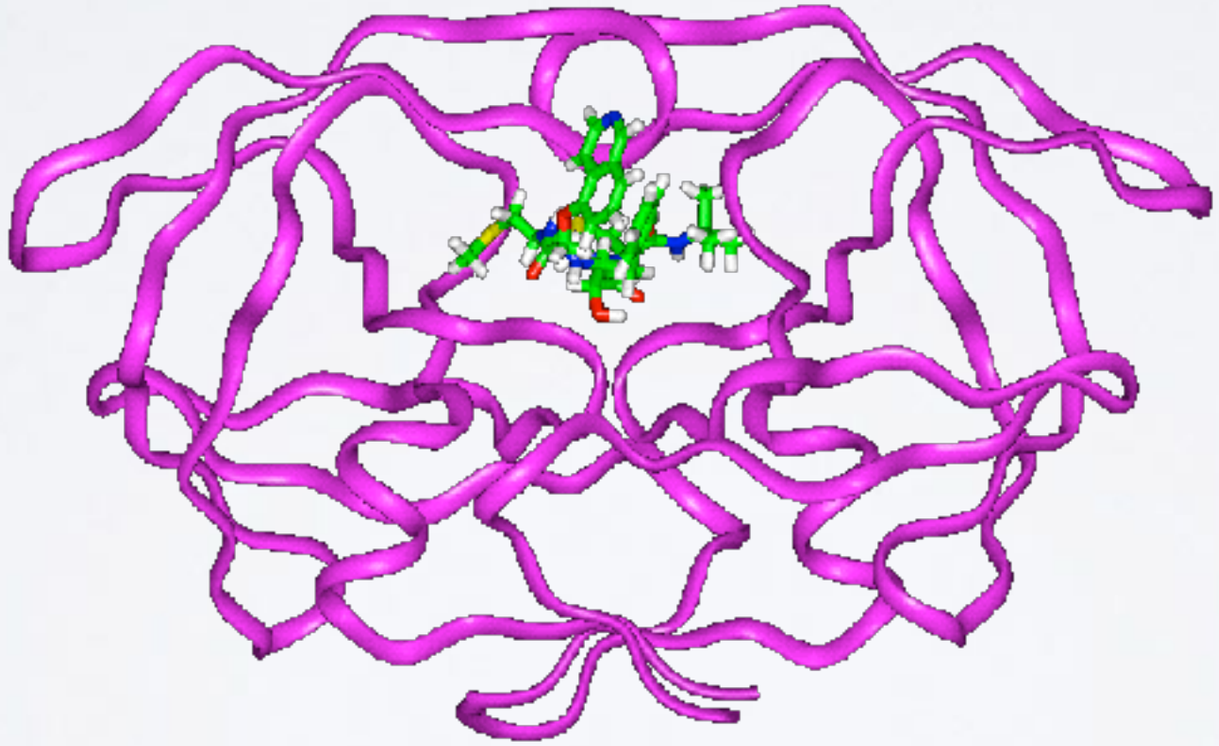
Two main approaches:

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

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

SCENARIO I:
RECEPTOR-BASED DRUG DISCOVERY

Structure of Targeted Protein Known: Structure-Based Drug Discovery



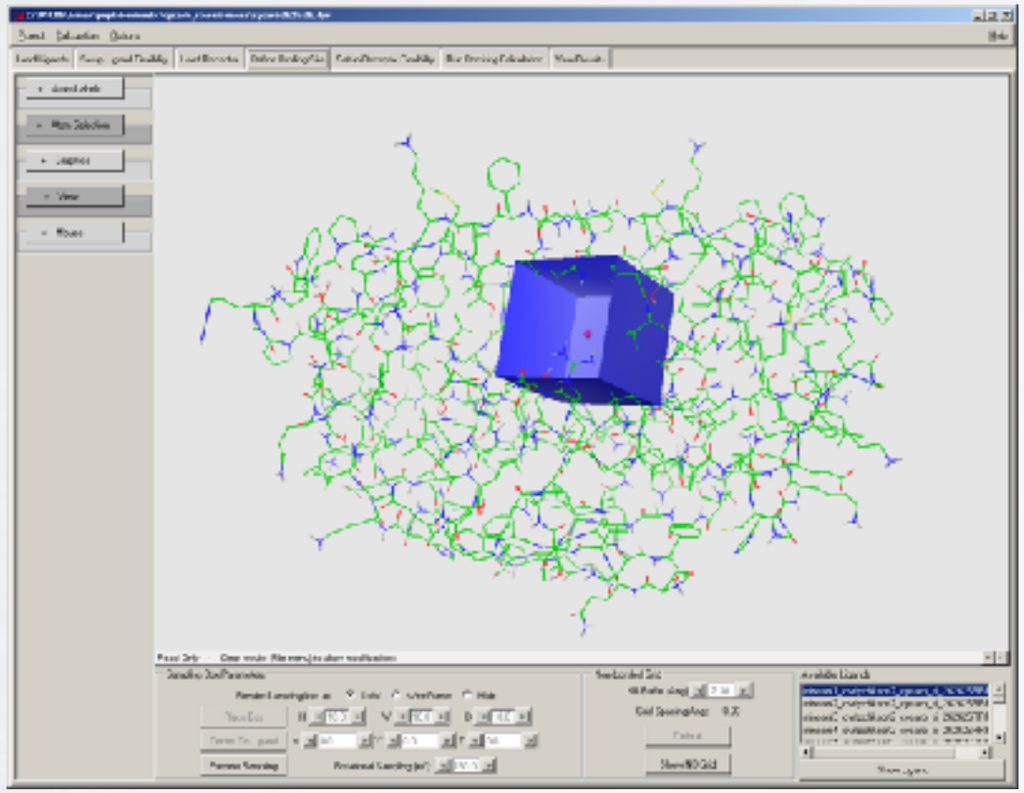
HIV Protease/KNI-272 complex

PROTEIN-LIGAND DOCKING

Structure-Based Ligand Design

Docking software

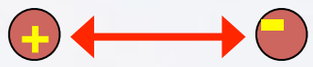
Search for structure of lowest energy



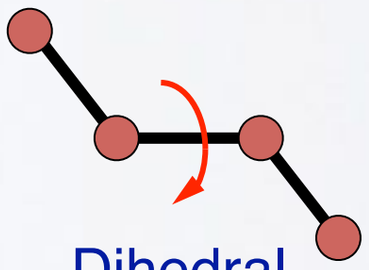
Potential function

Energy as function of structure

VDW

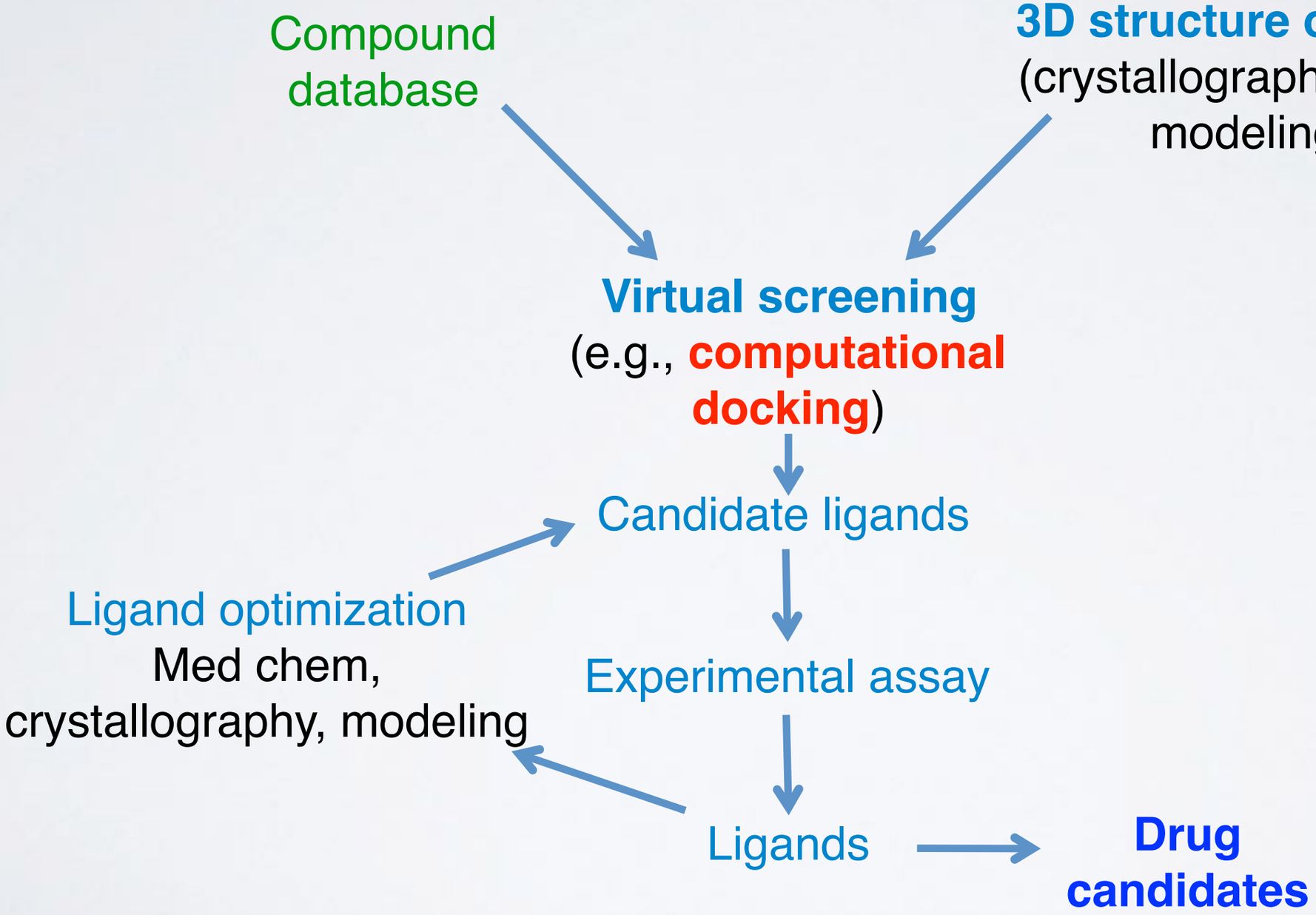


Screened Coulombic



Dihedral

STRUCTURE-BASED VIRTUAL SCREENING



COMPOUND LIBRARIES

Commercial
(in-house pharma)

Government (NIH)

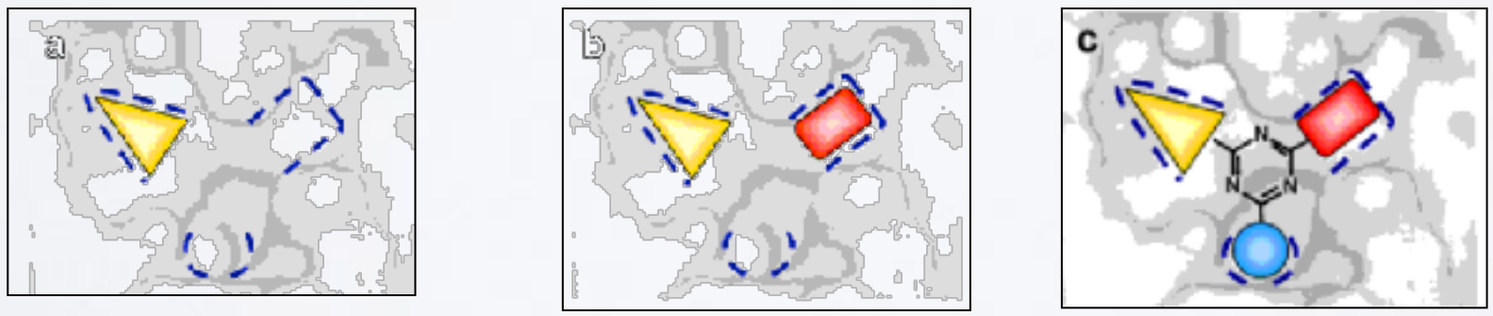
Academia

FRAGMENTAL STRUCTURE-BASED SCREENING

“Fragment” library → **Fragment docking** ← **3D structure of target**

Fragment docking → Compound design → Experimental assay and ligand optimization → **Drug candidates**

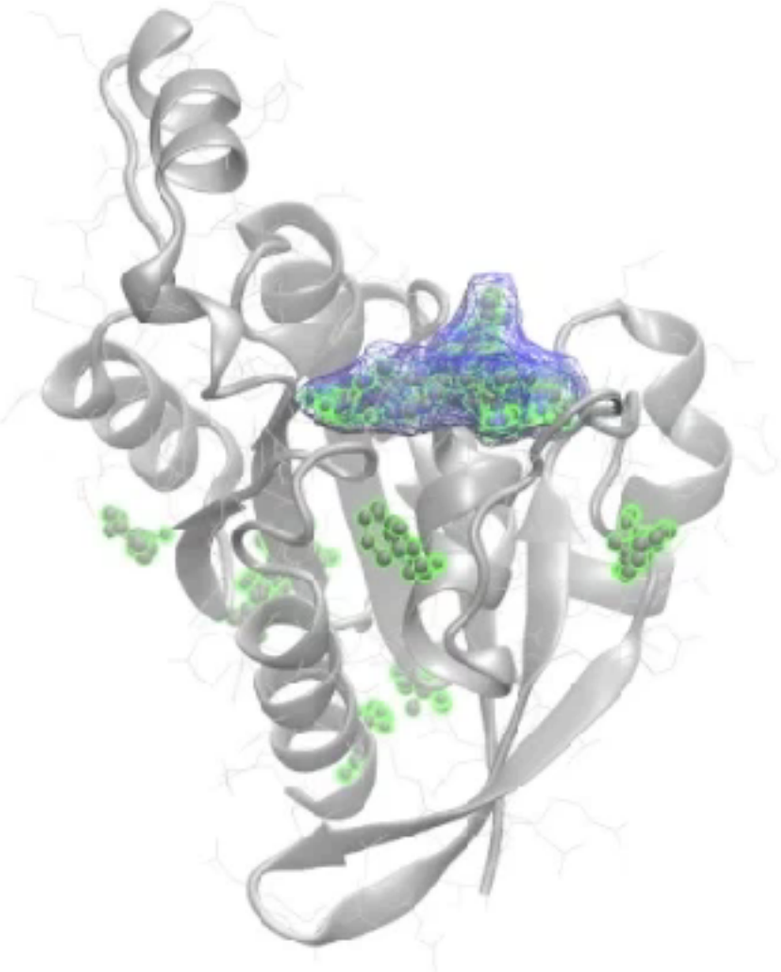
Med chem, crystallography, modeling

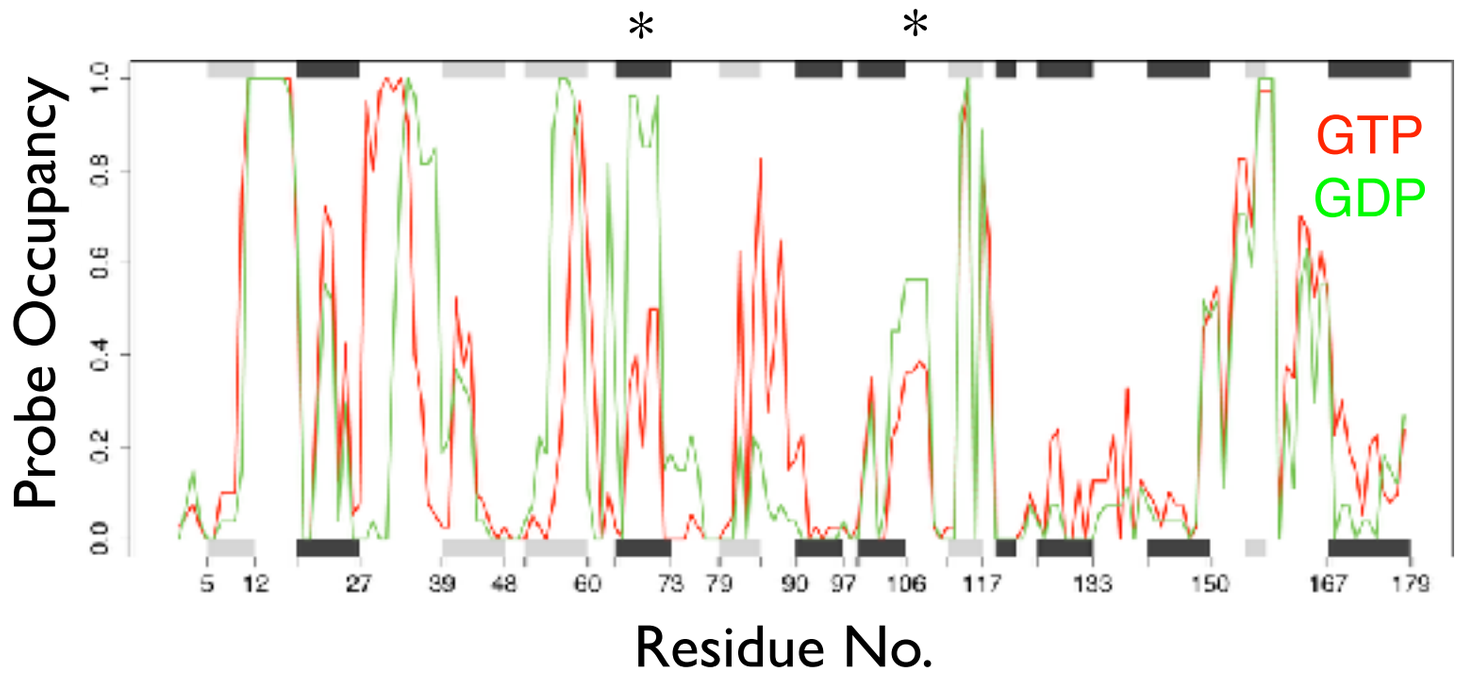


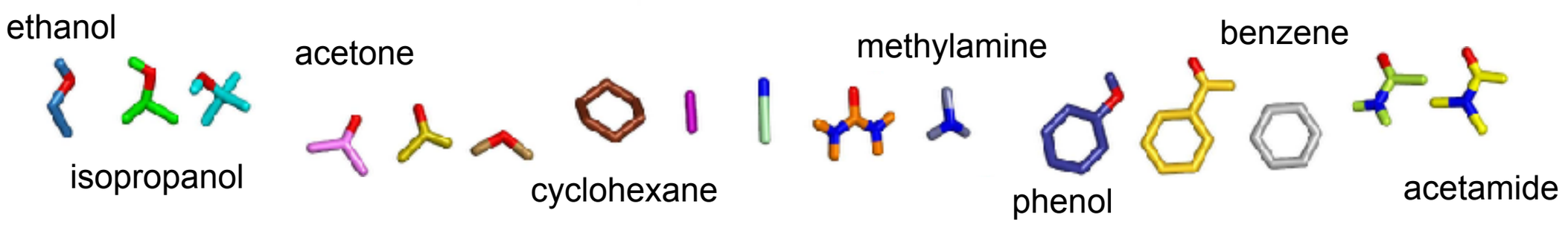
http://www.beilstein-institut.de/bozen2002/proceedings/Jhoti/jhoti.html

Multiple non active-site pockets identified

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



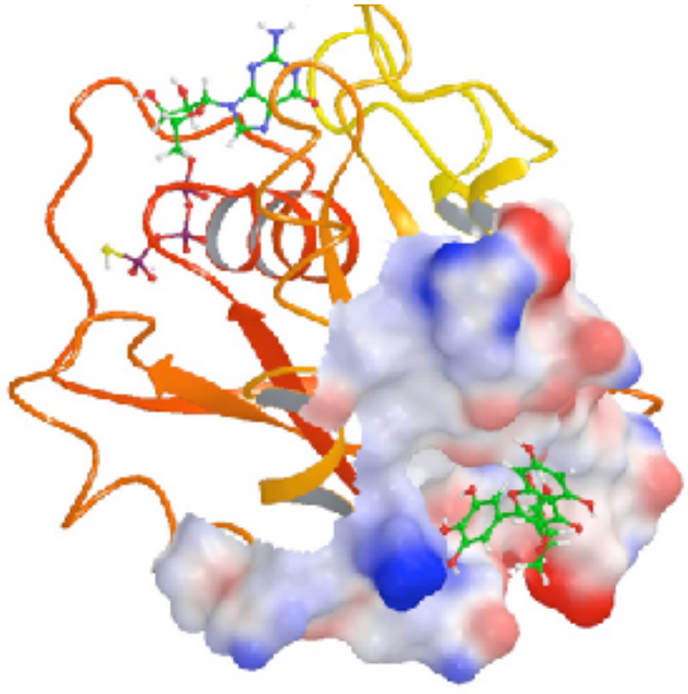


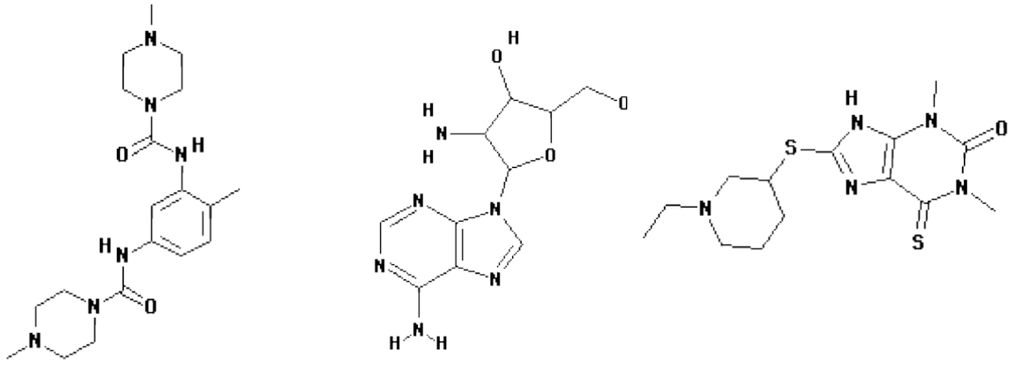


Ensemble docking & candidate inhibitor testing

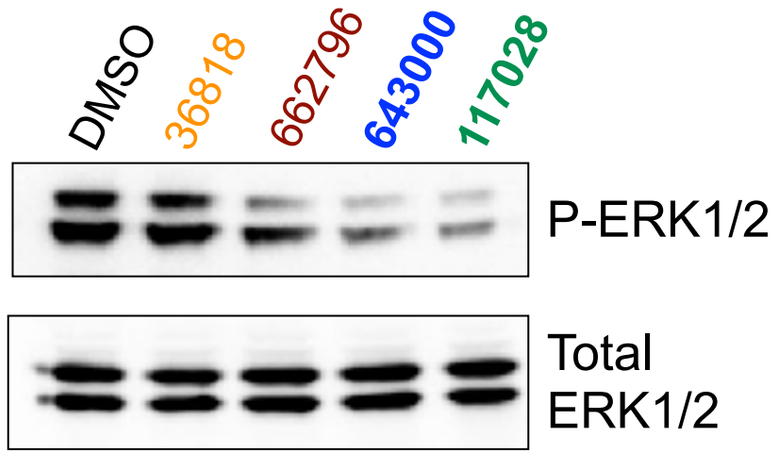
Top hits from ensemble docking against distal pockets were tested for inhibitory effects on basal ERK activity in glioblastoma cell lines.

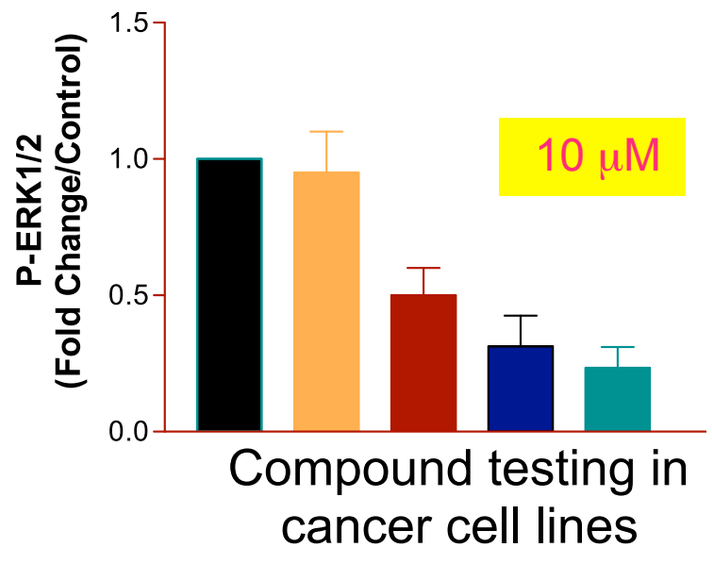
Ensemble computational docking





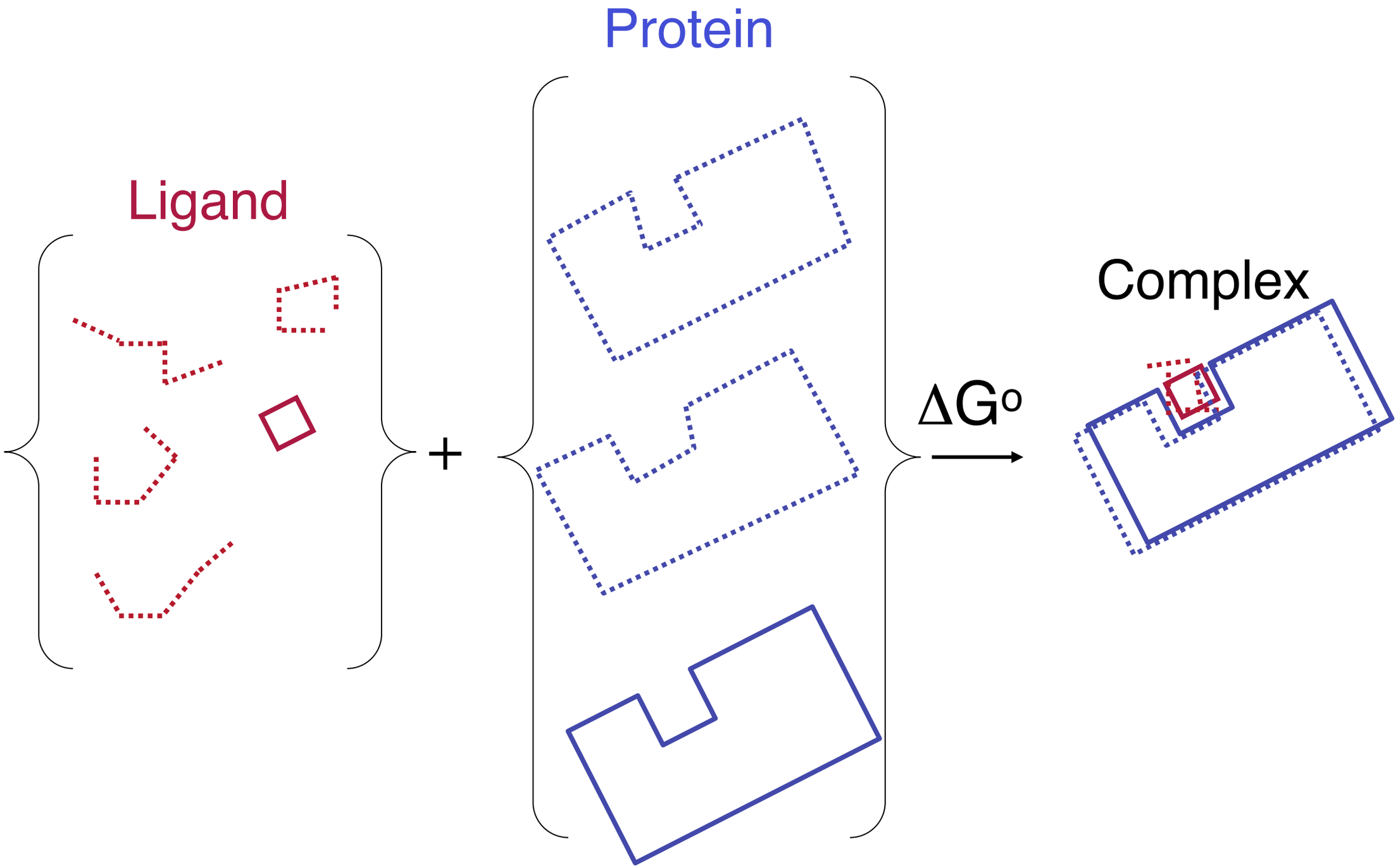
Compound effect on U251 cell line





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**

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

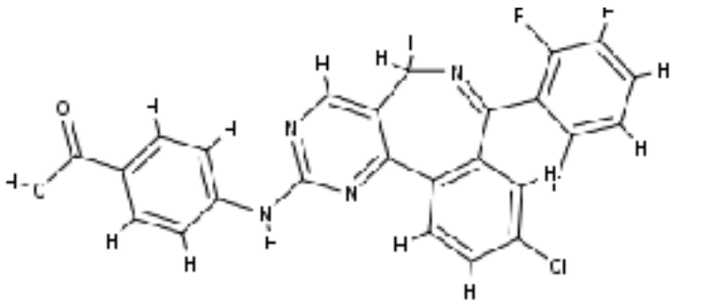
A company wants to work around another company's chemical patents

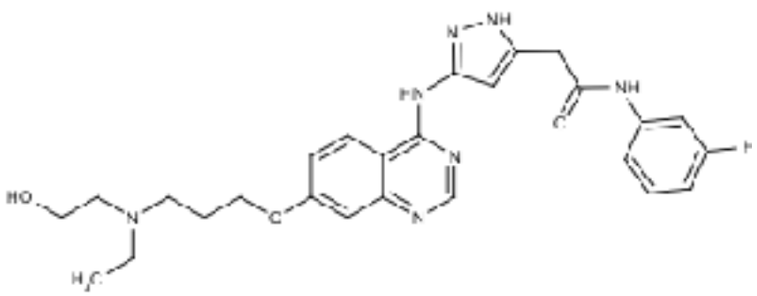
A high-affinity ligand is toxic, is not well-absorbed, etc.

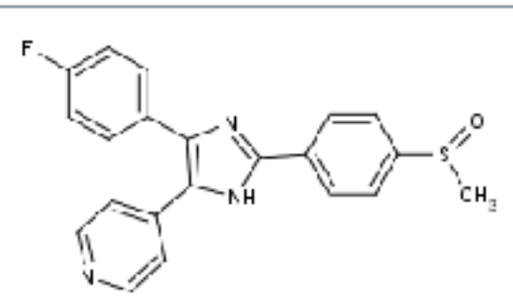
Scenario 2

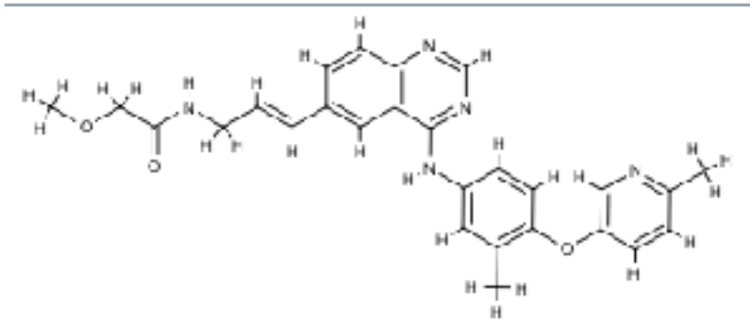
Structure of Targeted Protein Unknown: Ligand-Based Drug Discovery

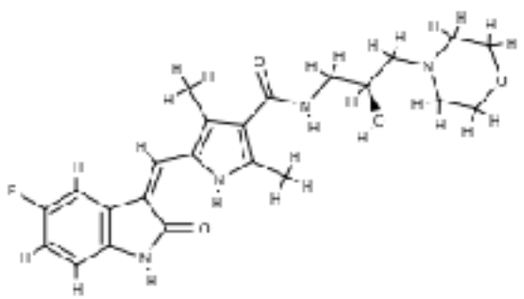
e.g. MAP Kinase Inhibitors

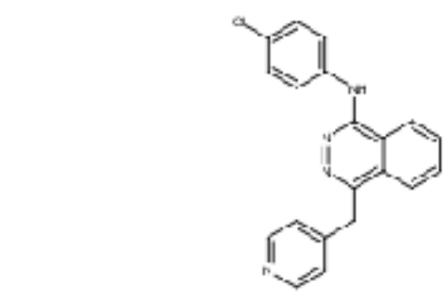


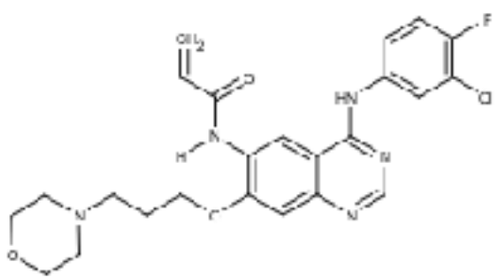


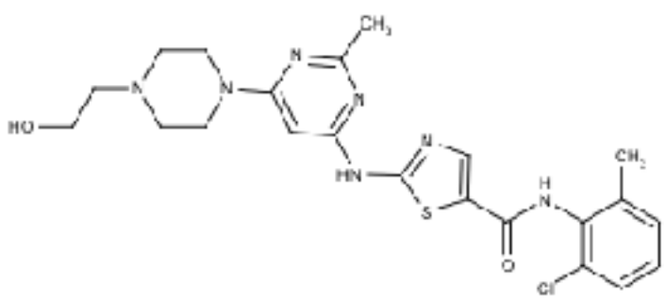












Using knowledge of existing inhibitors to discover more

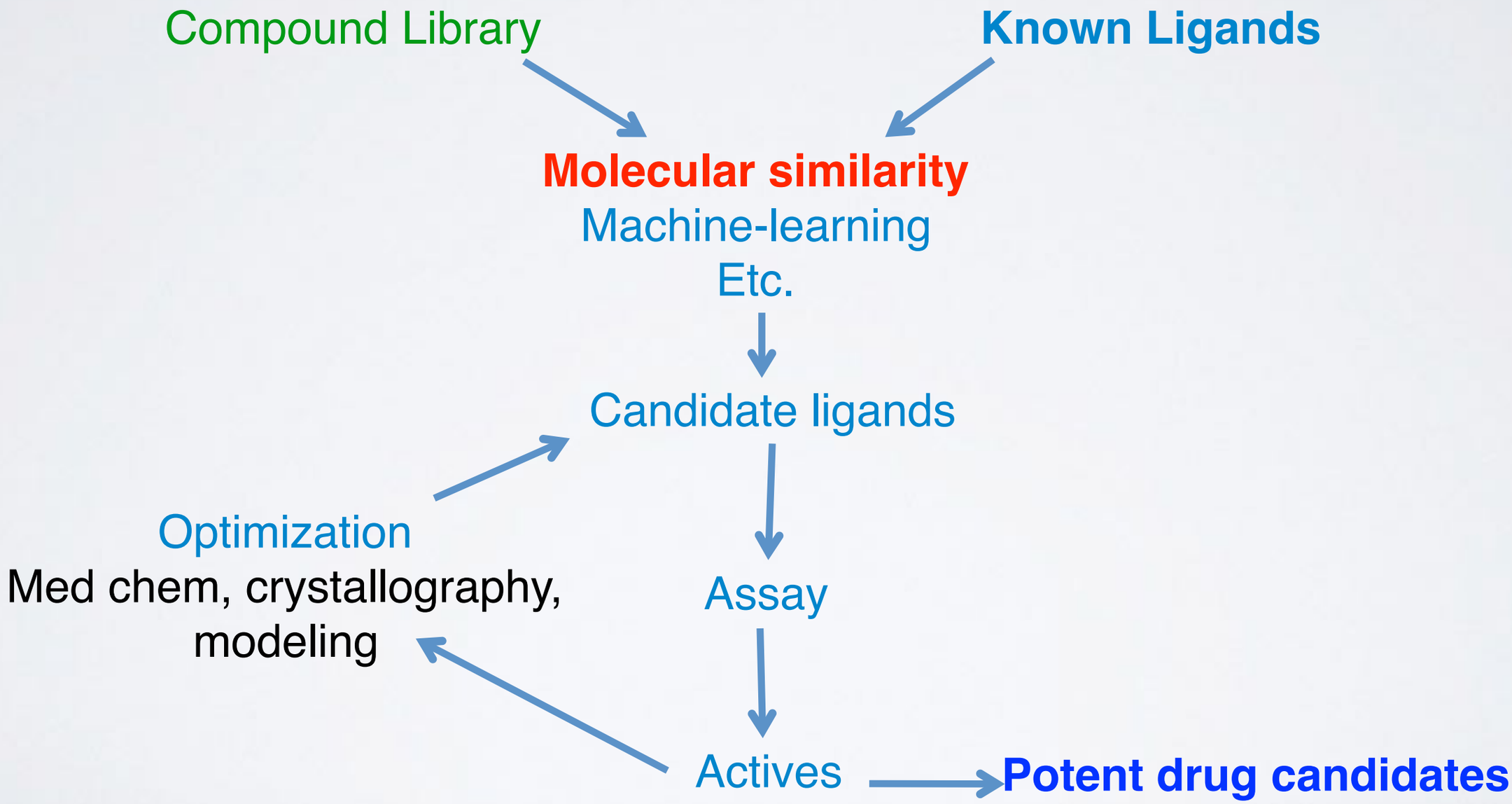
Why Look for Another Ligand if You Already Have Some?

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

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

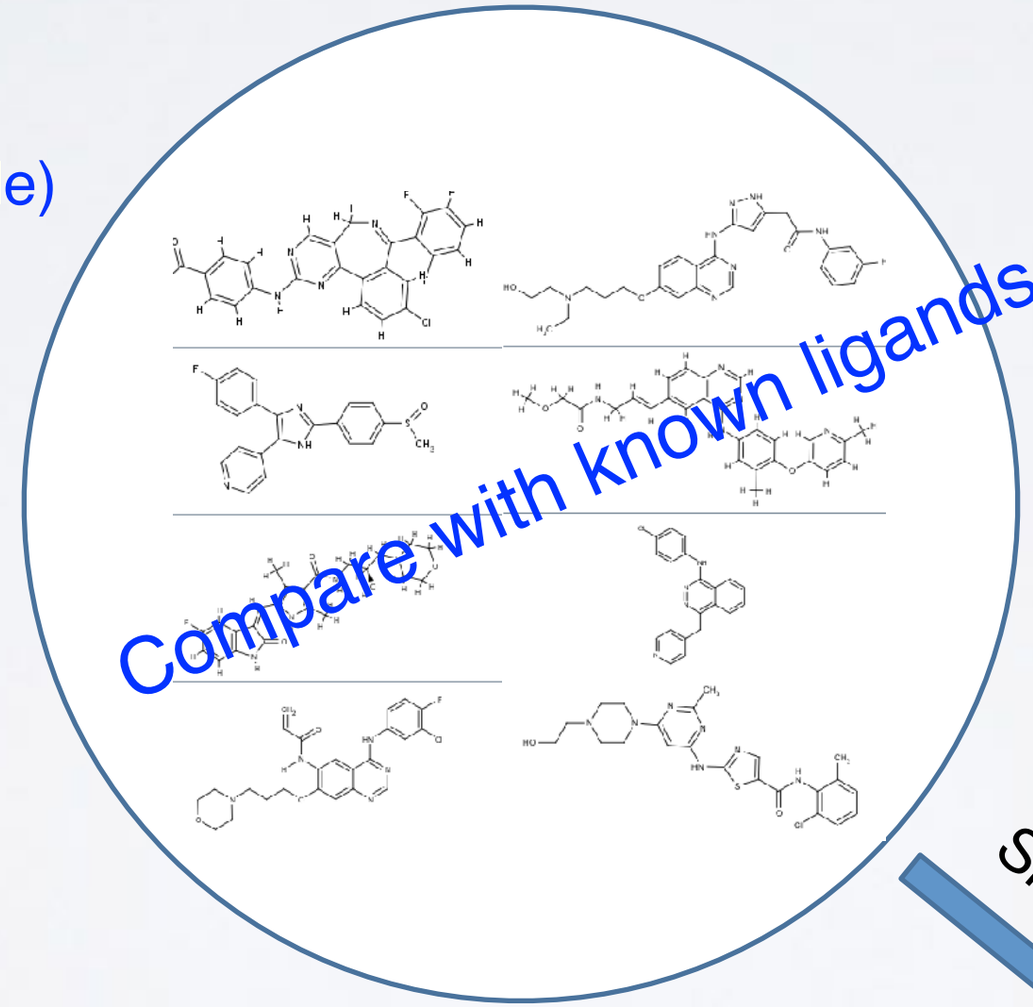
An high-affinity ligand is toxic, is not well-absorbed, etc.

LIGAND-BASED VIRTUAL SCREENING



CHEMICAL SIMILARITY
LIGAND-BASED DRUG-DISCOVERY

Compounds
(available/synthesizable)



Different

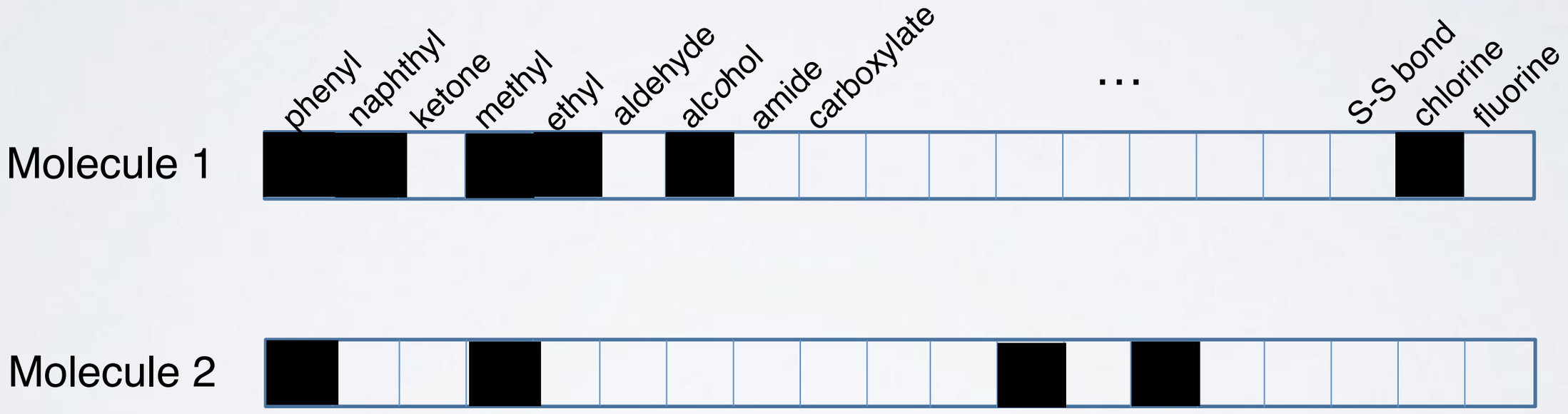
Don’t bother

Similar

Test experimentally

CHEMICAL FINGERPRINTS

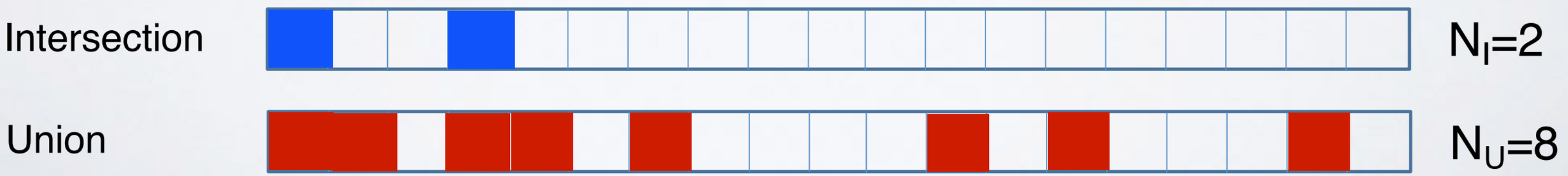
BINARY STRUCTURE KEYS



CHEMICAL SIMILARITY FROM FINGERPRINTS

Tanimoto Similarity (or Jaccard Index), T

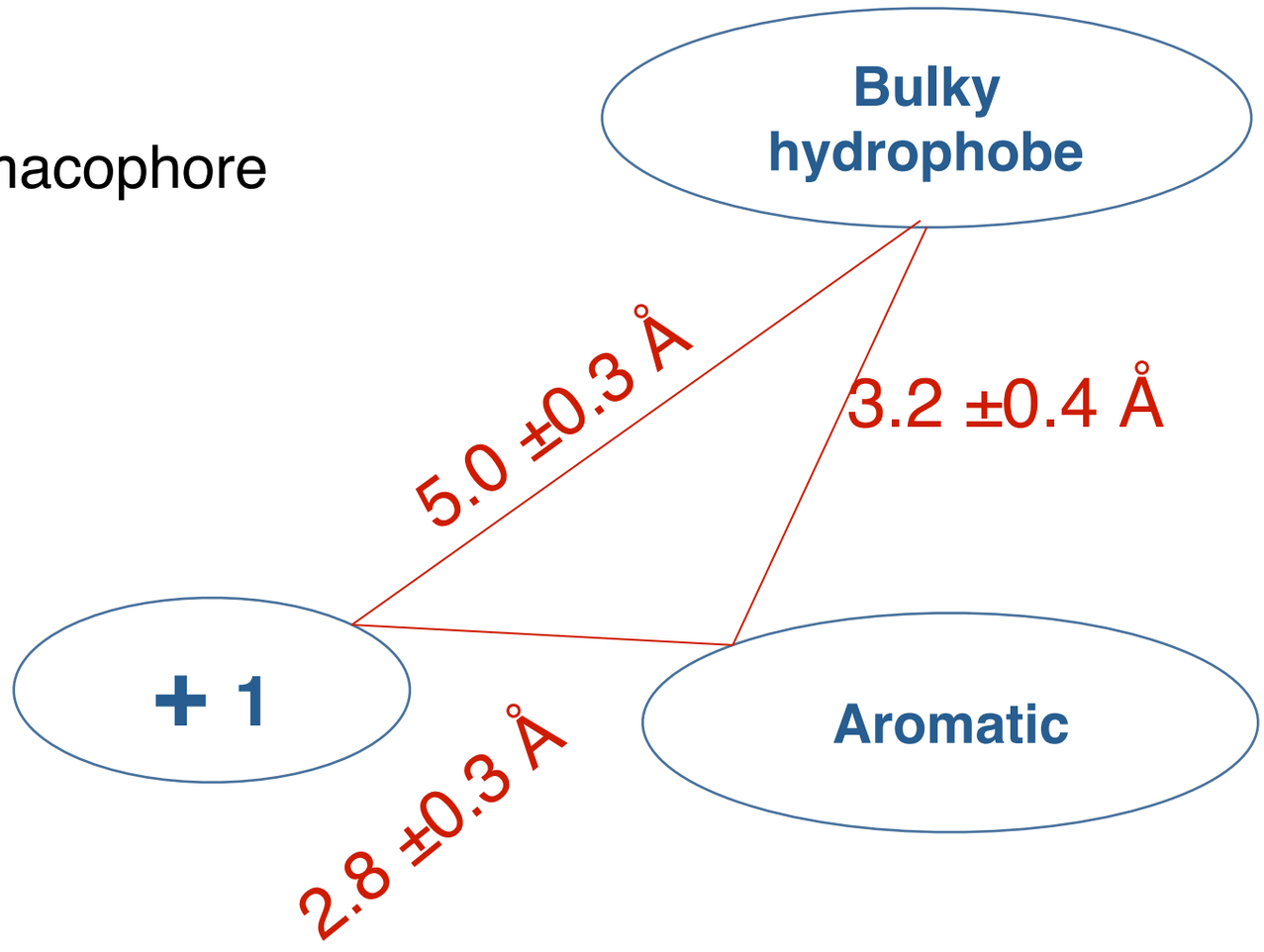
$$T \equiv \frac{N_I}{N_U} = 0.25$$



Pharmacophore Models

Φάρμακο (drug) + Φορά (carry)

A 3-point pharmacophore

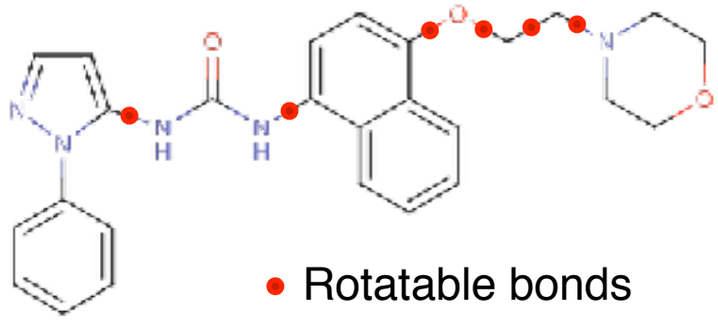


Molecular Descriptors

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

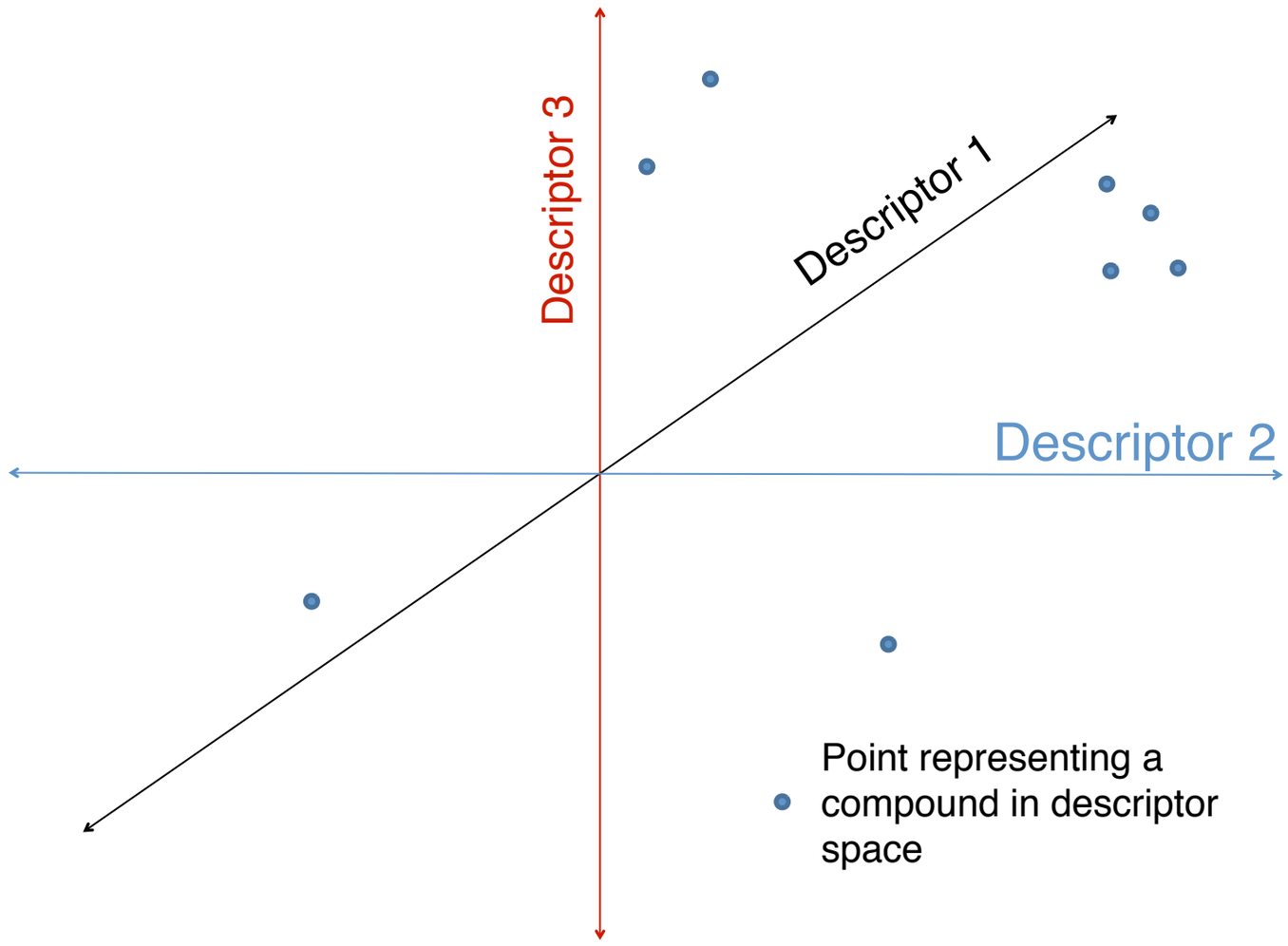
- branching index
- measures of linearity vs interconnectedness

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.

Do it Yourself!

Hand-on time!

https://bioboot.github.io/bggn213_f17/lectures/#12

You can use the classroom computers or your own laptops. If you are using your laptops then you will need to install **VMD** and **MGLTools**

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

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

NEXT UP:

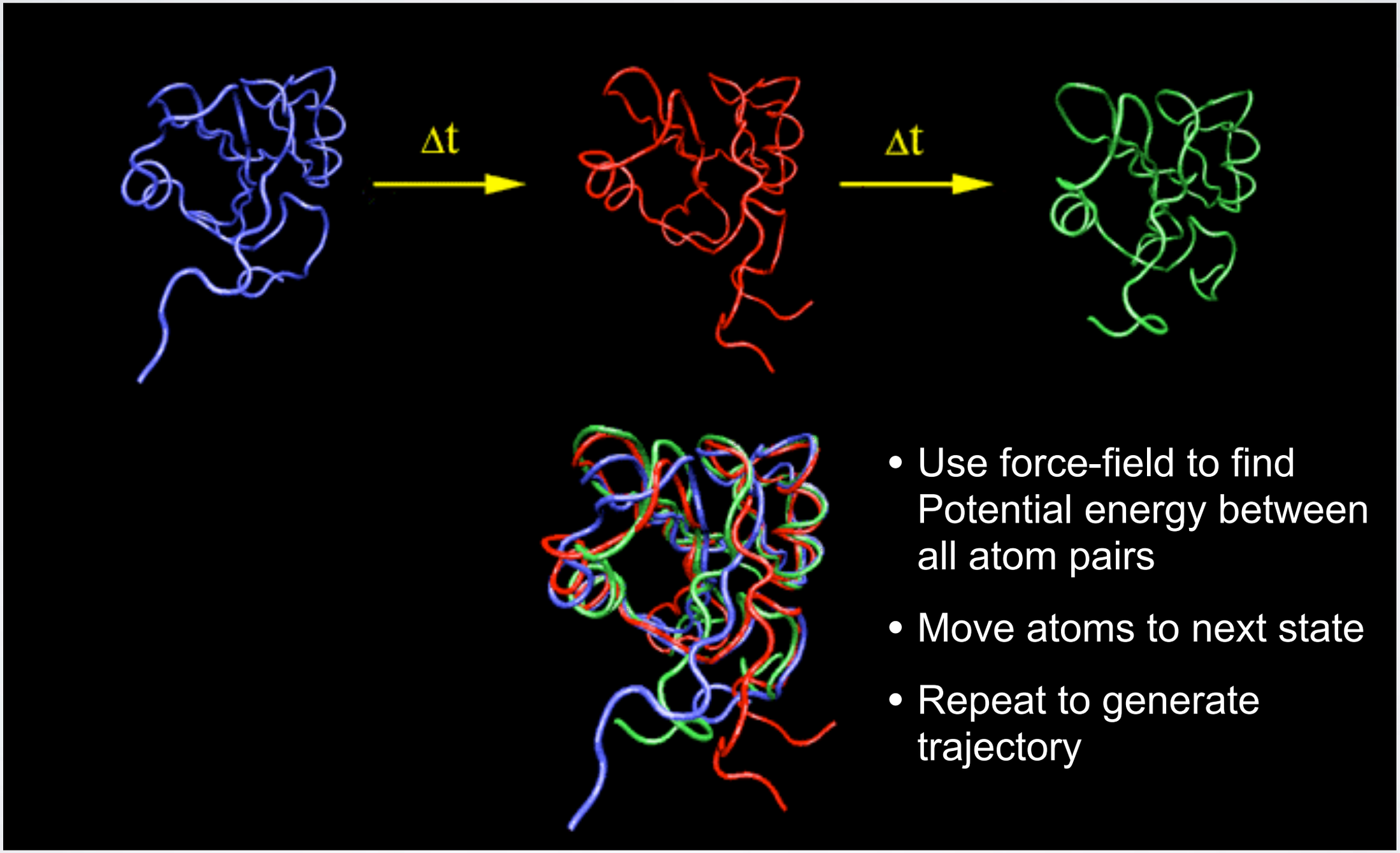
- **Overview of structural bioinformatics**
  - Major motivations, goals and challenges
- **Fundamentals of protein structure**
  - Composition, form, forces and dynamics
- **Representing and interpreting protein structure**
  - Modeling energy as a function of structure
- **Example application areas**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



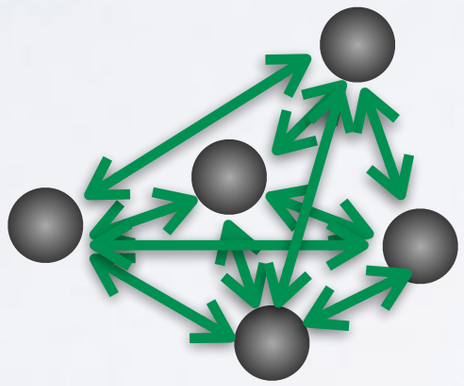
McCammon, Gelin & Karplus, *Nature* (1977)

[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

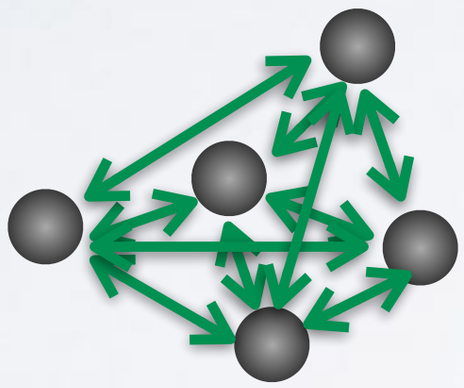
$$m_i \frac{d^2}{dt^2} \vec{R}_i = -\vec{\nabla}_i E(\vec{R})$$

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)



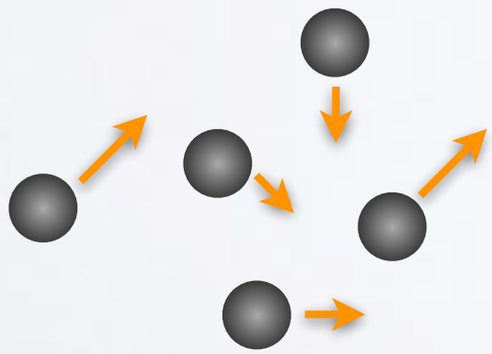
Nucleic motion described classically

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

Empirical force field

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

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



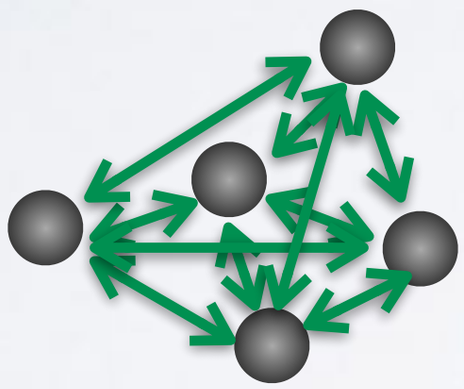
$$\boldsymbol{v}(t + \frac{\Delta t}{2}) = \boldsymbol{v}(t - \frac{\Delta t}{2}) + \frac{\boldsymbol{F}(t)}{m} \Delta t$$

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

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

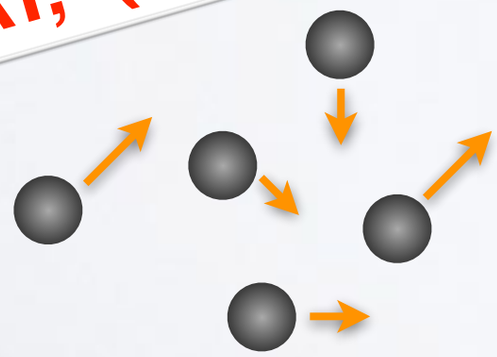
$$m_i \frac{d^2}{dt^2}\vec{R}_i = -\vec{\nabla}_i E(\vec{R})$$

Empirical force fi

$$E(\vec{R}) = \sum \ldots E_i(\vec{R})$$

…ion – bonded

- Use the fo… …date **velocities** and move atoms to new **positions**
  (b… …g numerically via the “leapfrog” scheme)

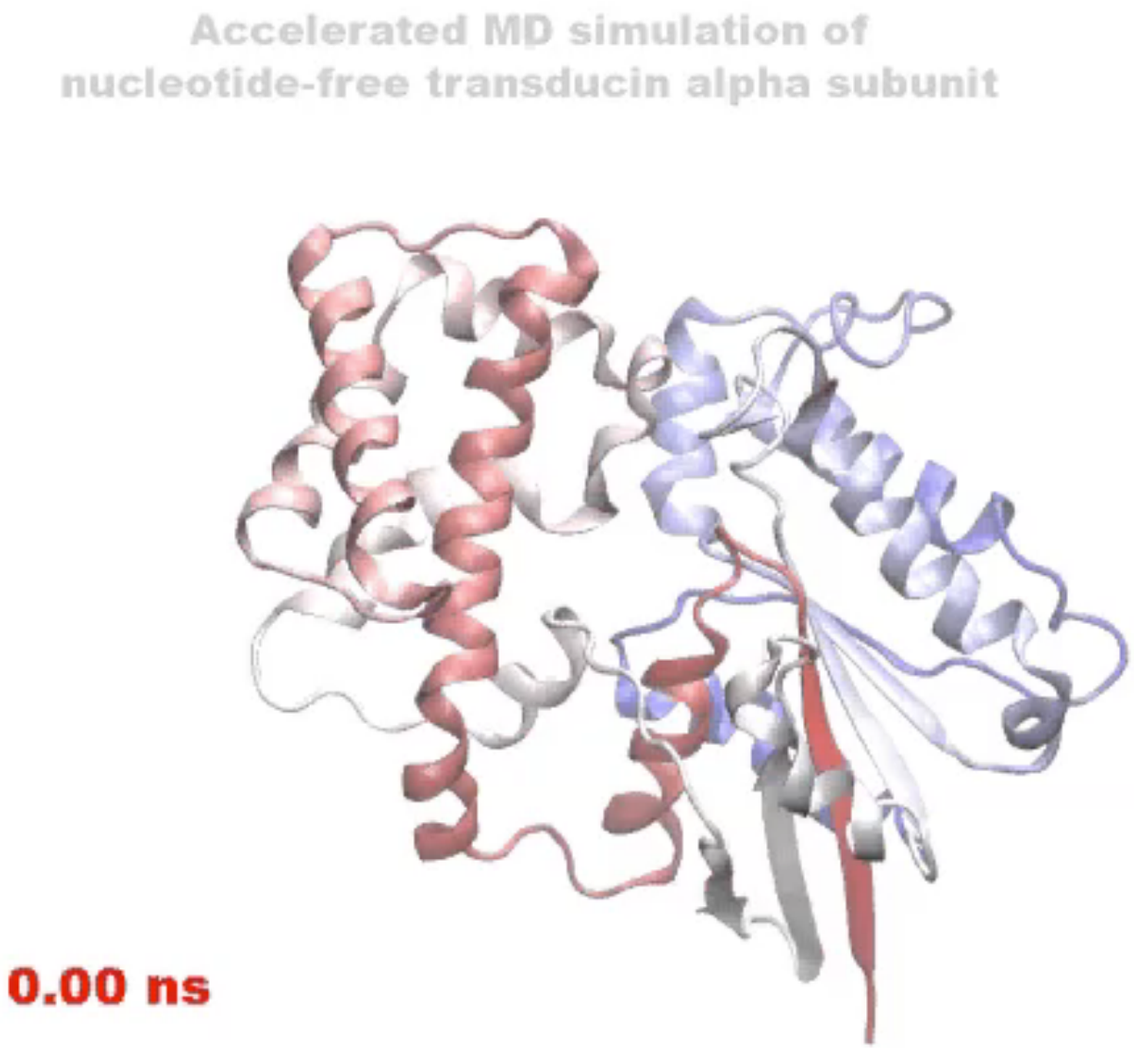


$$\boldsymbol{v}(t + \frac{\Delta t}{2}) = \boldsymbol{v}(t - \frac{\Delta t}{2}) + \frac{\boldsymbol{F}(t)}{m}\Delta t$$

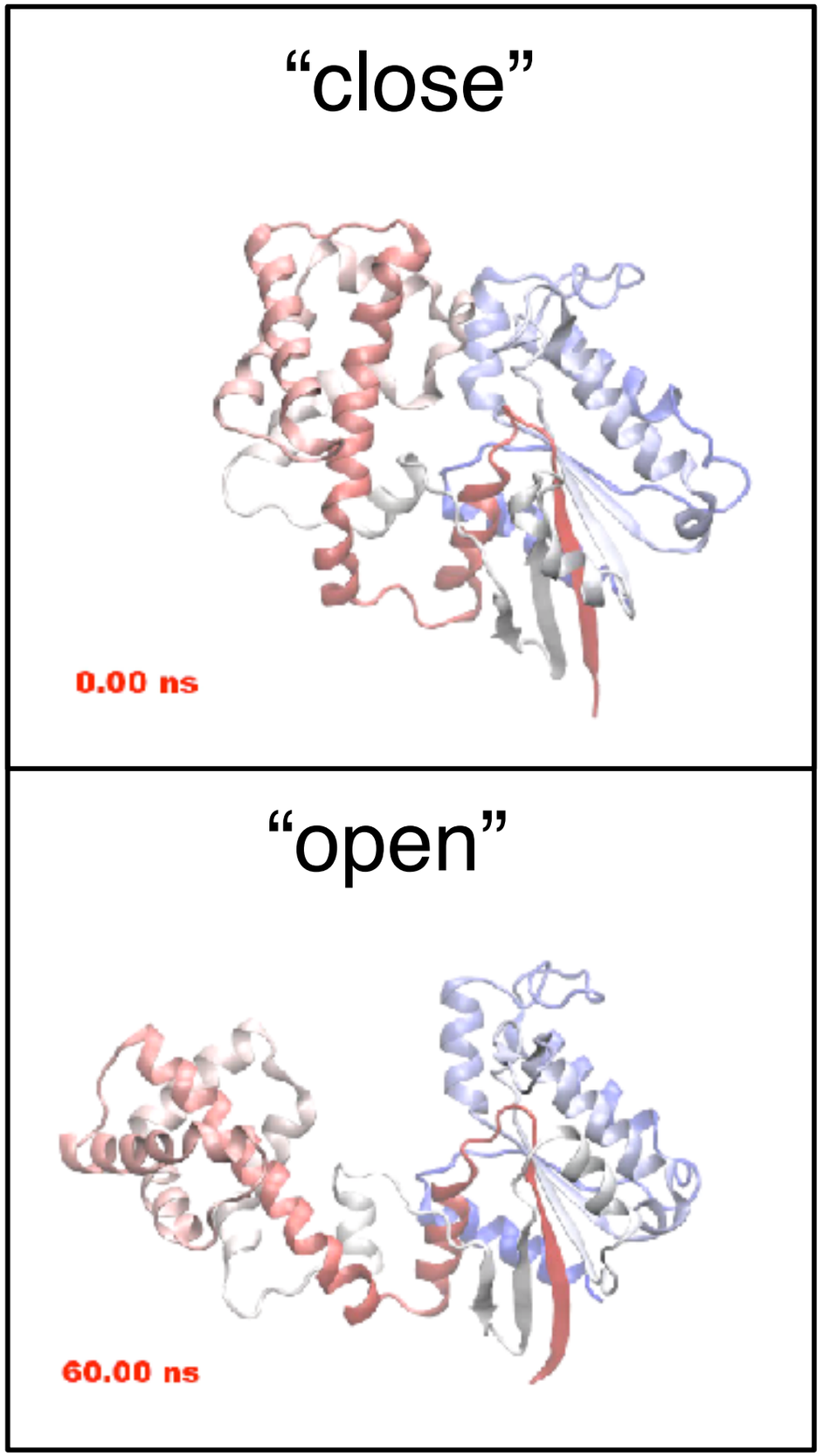
$$\boldsymbol{r}(t + \Delta t) = \boldsymbol{r}(t) + \boldsymbol{v}(t + \frac{\Delta t}{2})\Delta t$$

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

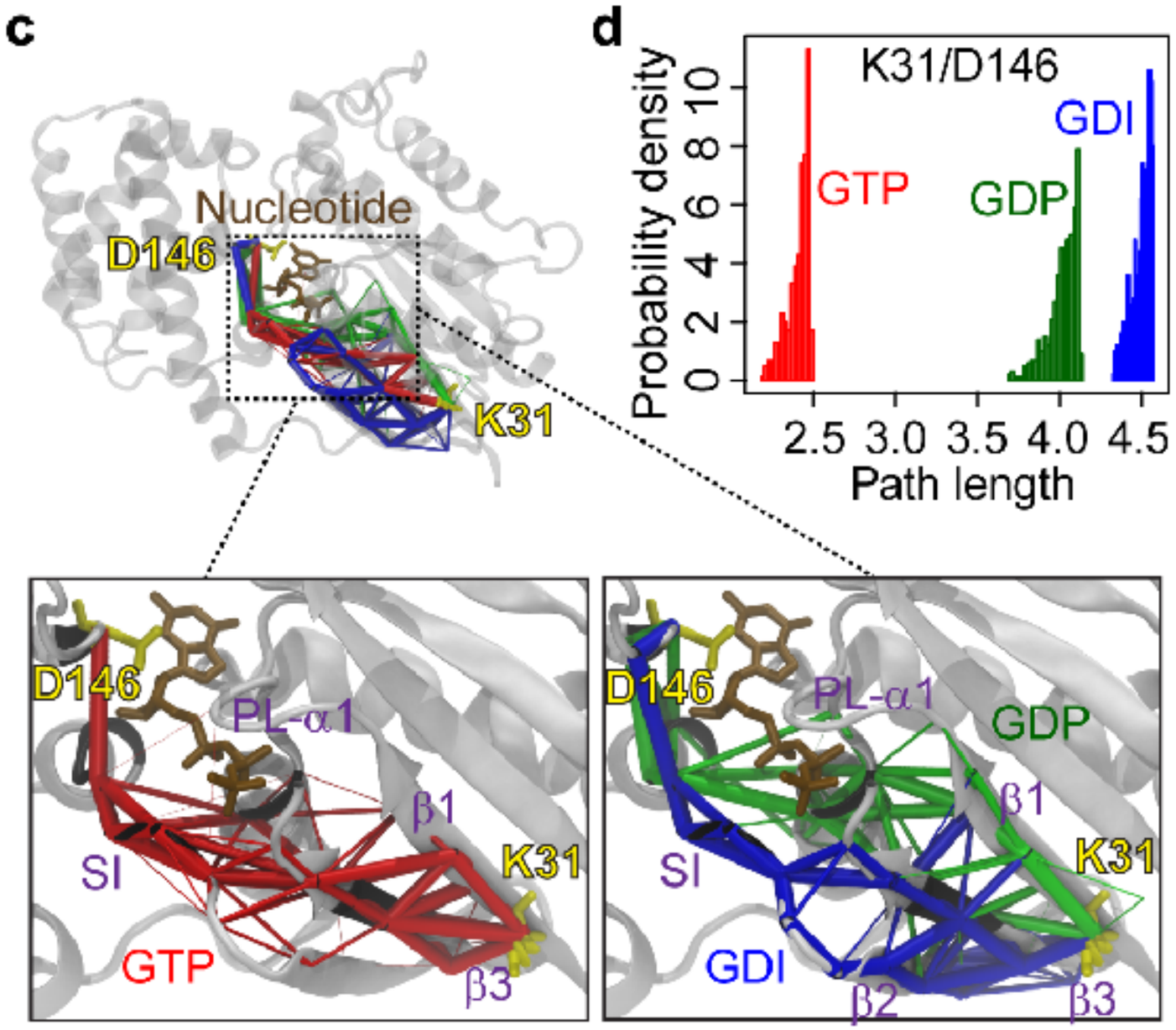
MD Prediction of Functional Motions



Yao and Grant, Biophys J. (2013)

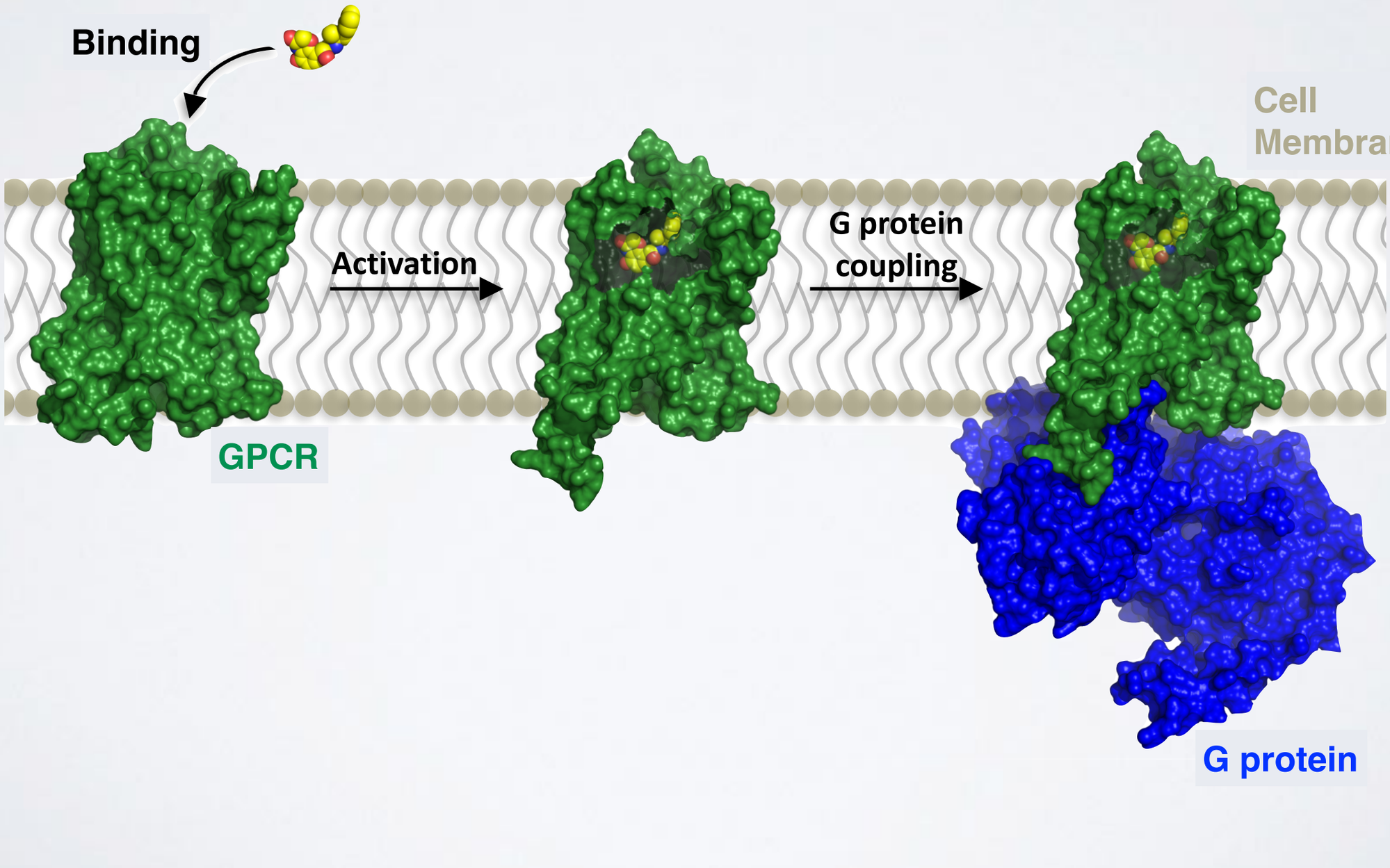


Simulations Identify Key Residues Mediating Dynamic Activation

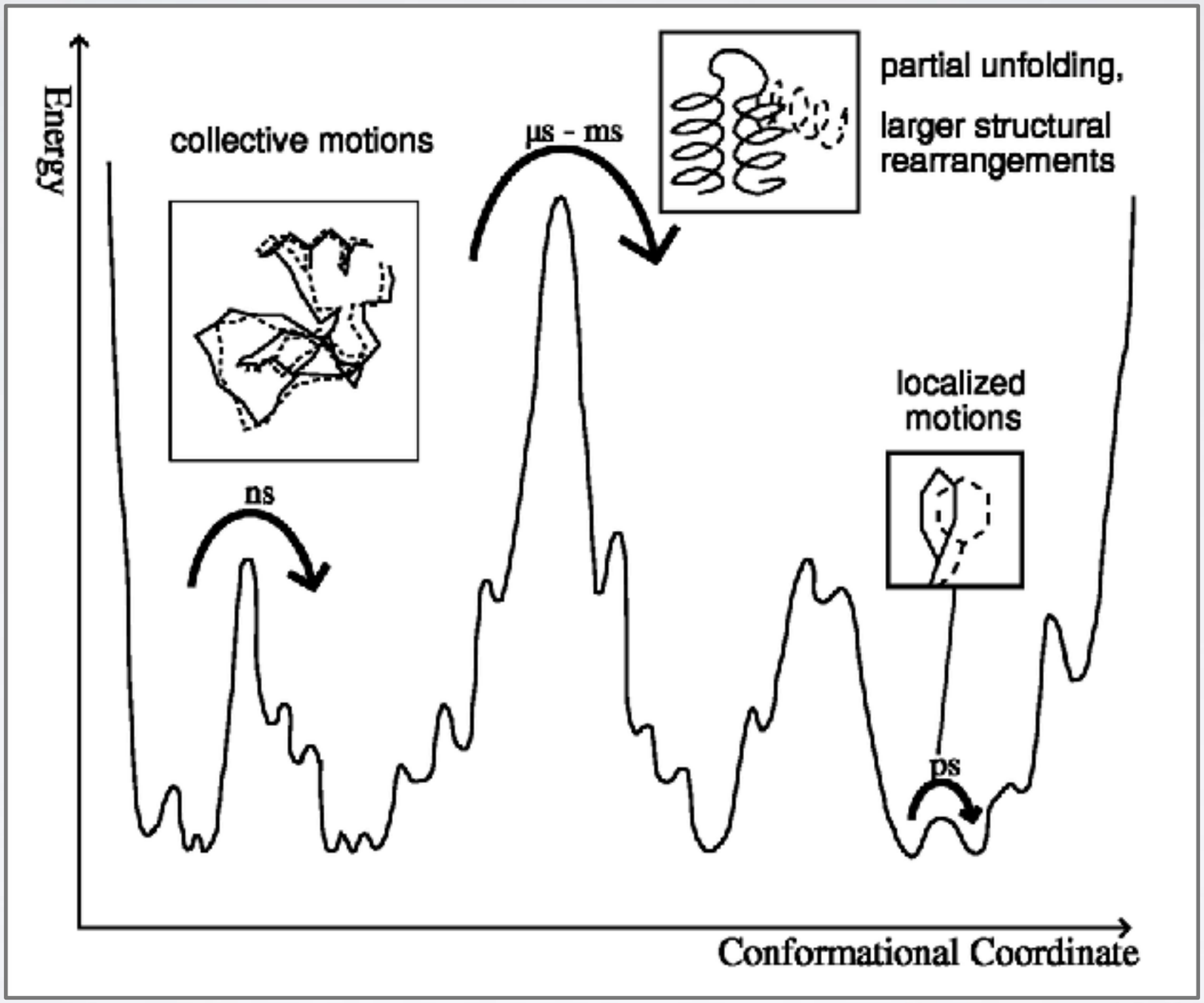


Yao … Grant, Journal of Biological Chemistry (2016)

EXAMPLE APPLICATION OF MOLECULAR SIMULATIONS TO GPCRS



PROTEINS JUMP BETWEEN MANY, HIERARCHICALLY ORDERED "CONFORMATIONAL SUBSTATES"



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

MOLECULAR DYNAMICS IS VERY

Improve this slide

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

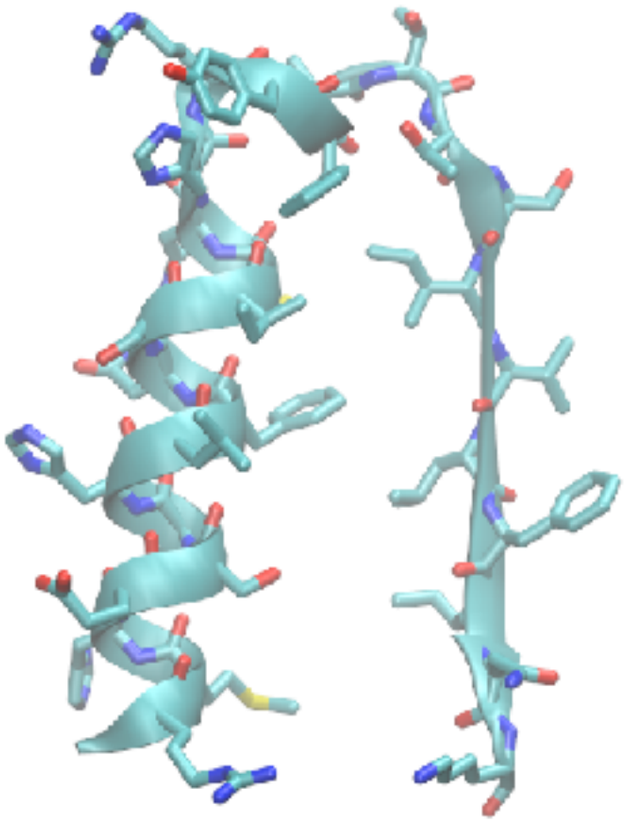
(on a 100 Gflop/s cpu: **ca 25 years!**)

… 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)**

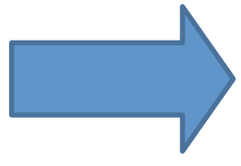
COARSE GRAINING: **NORMAL MODE ANALYSIS** (NMA)

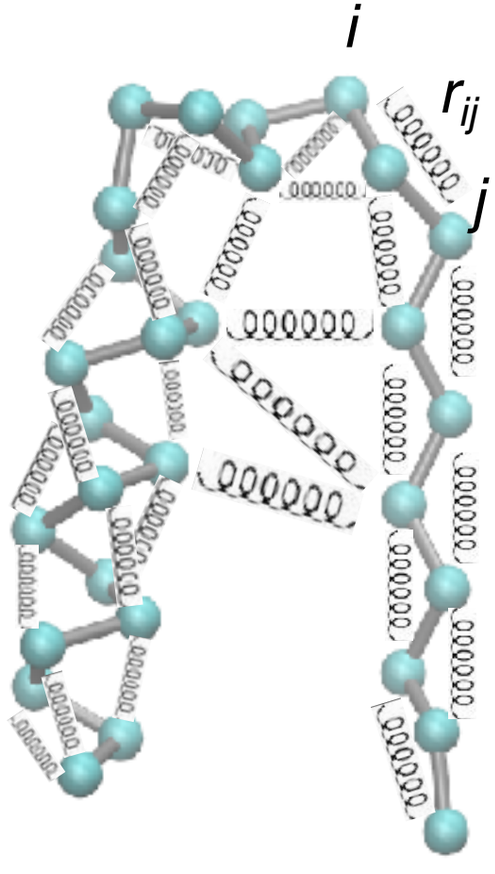
- MD is still time-consuming for large systems
- Elastic network model NMA (ENM-NMA) is an example of a lower resolution approach that finishes in seconds even for large systems.



Atomistic

C. G.

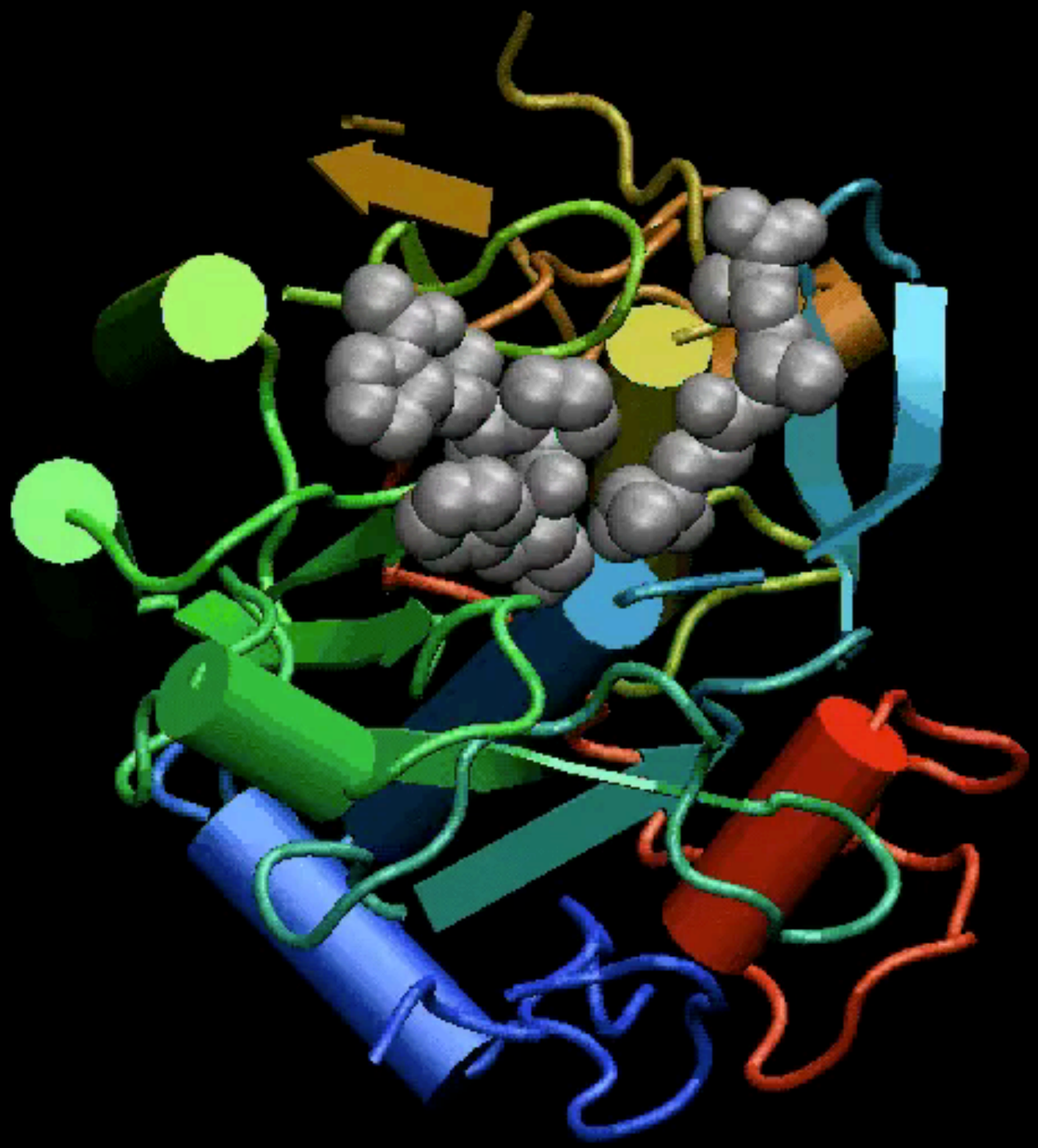




Coarse Grained

- 1 bead / 1 amino acid
- Connected by springs

NMA models the protein as a network of elastic strings



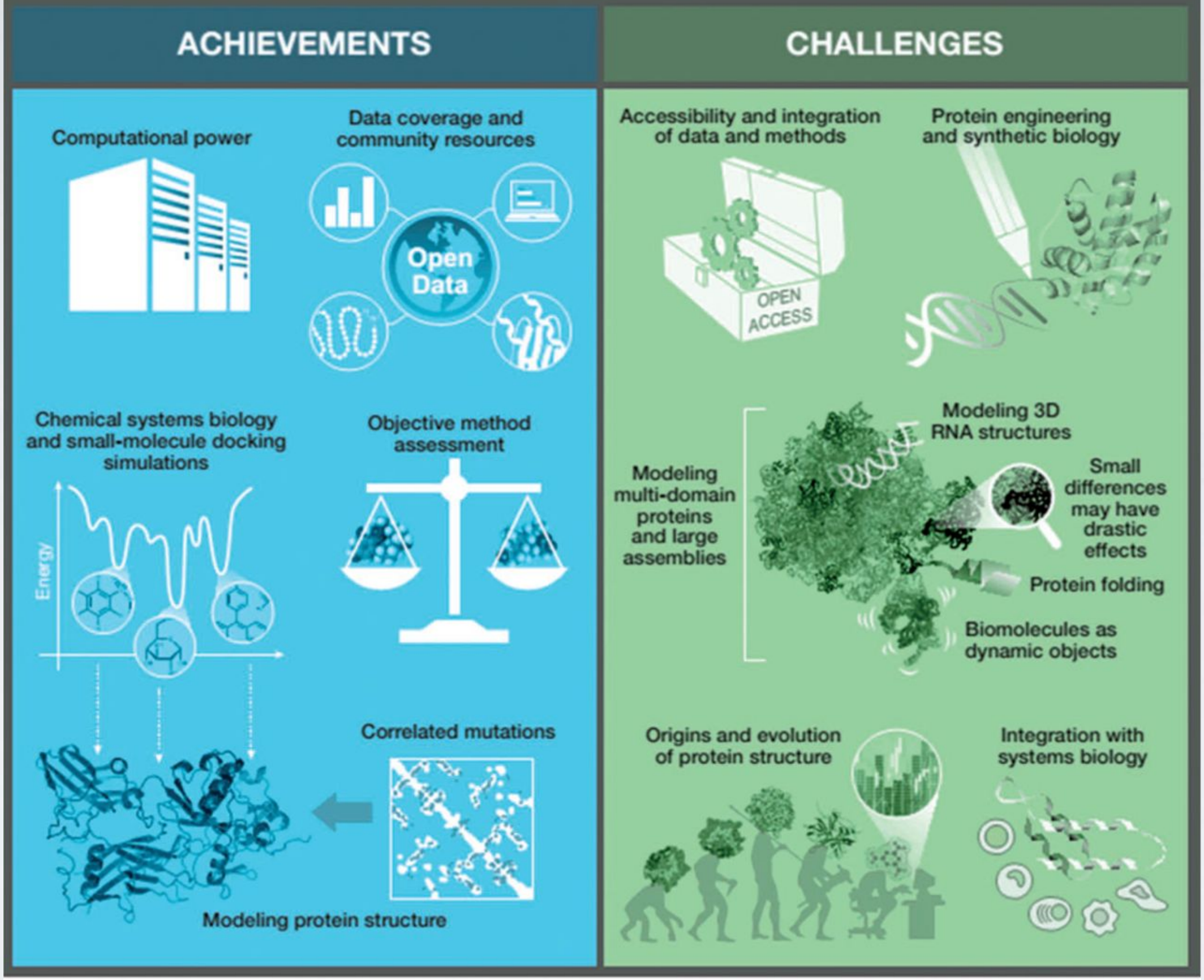
Proteinase K

Do it Yourself!

Hand-on time!

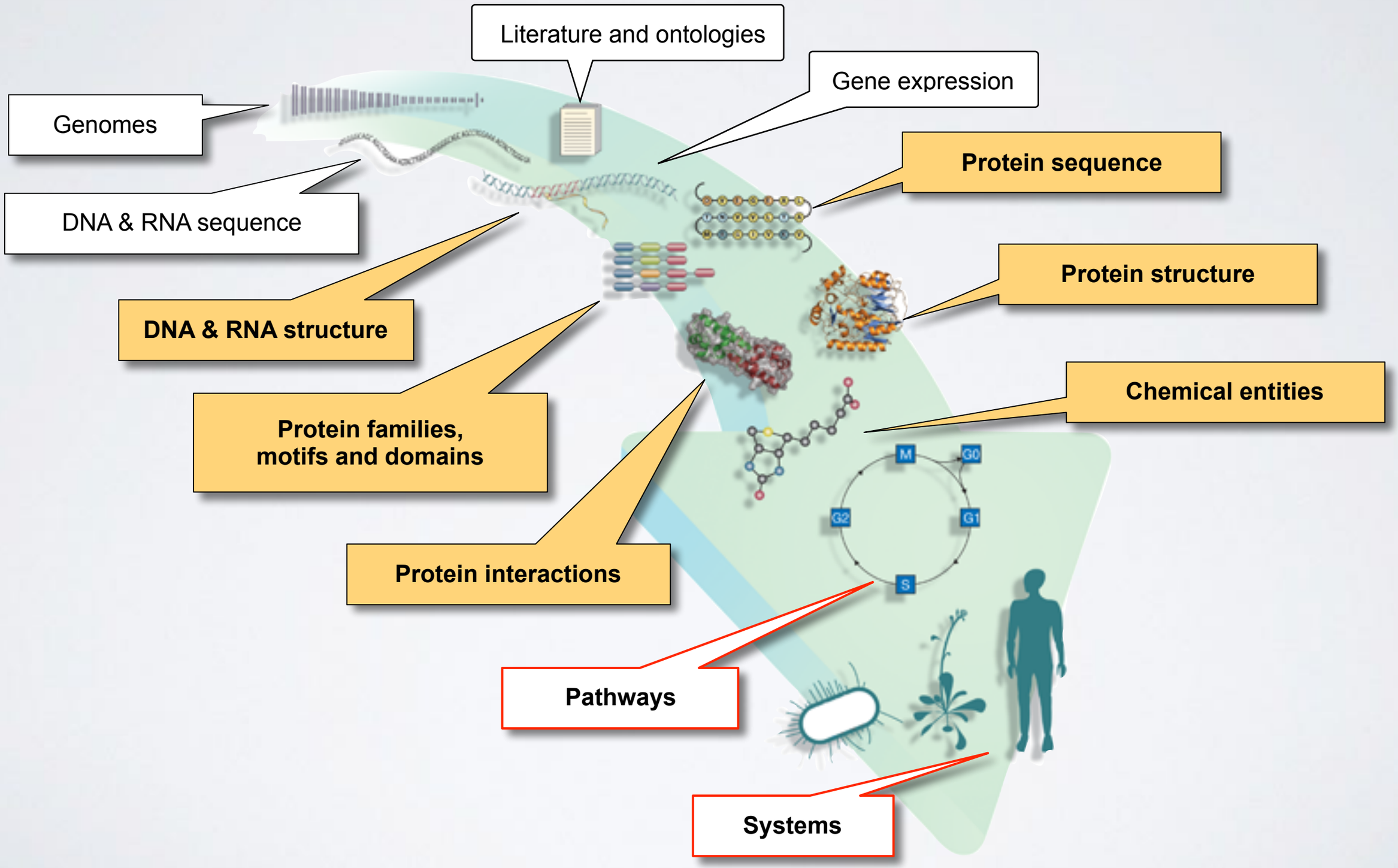
https://bioboot.github.io/bggn213_f17/lectures/#12

Focus on **section 4** exploring **PCA** and **NMA apps**



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

INFORMING SYSTEMS BIOLOGY?



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