



# BGGN 213

## Structural Bioinformatics II

Lecture 12

Barry Grant

UC San Diego

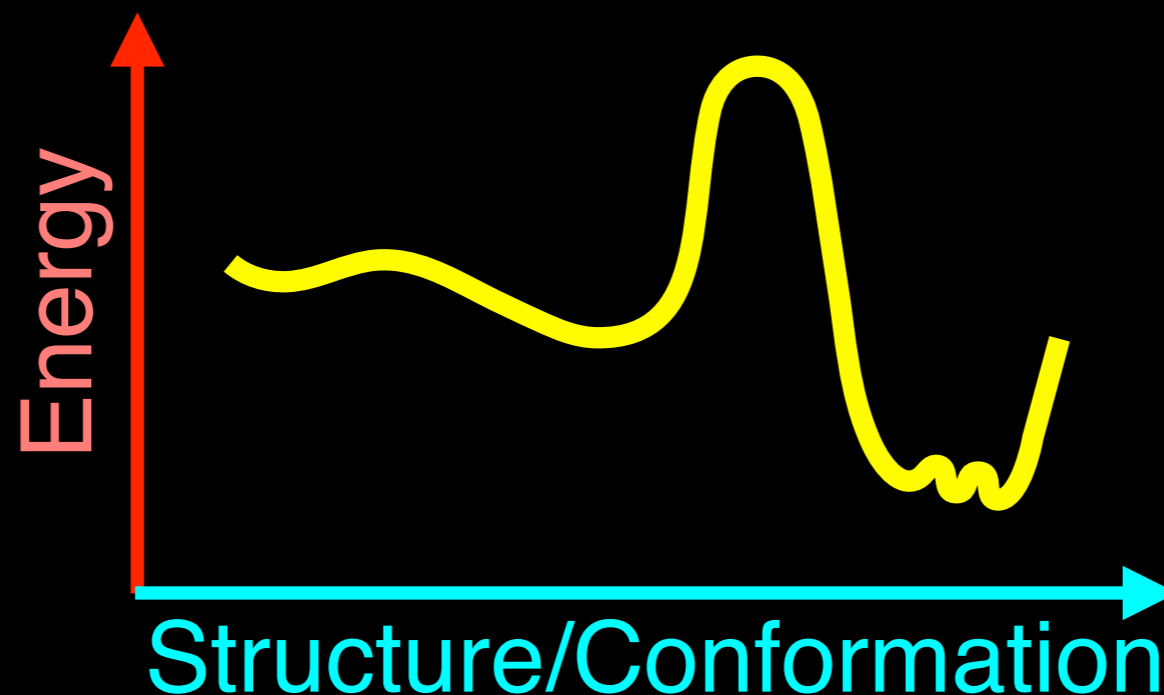
<http://thegrantlab.org/bggn213>

# Next Up:

- **Overview of structural bioinformatics**
  - Motivations, goals and challenges
- **Fundamentals of protein structure**
  - Structure composition, form and forces
- **Representing, interpreting & modeling protein structure**
  - Visualizing and interpreting protein structures
  - Analyzing protein structures
  - Modeling energy as a function of structure
  - Drug discovery & Predicting functional dynamics

# Key concept:

Potential functions describe a system's energy as a function of its structure



Two main approaches:

**(1). Physics-Based**

**(2). Knowledge-Based**

Two main approaches:

**(1). Physics-Based**

**(2). Knowledge-Based**

For **physics** based potentials  
energy terms come from physical theory

$$V(R) = E_{\text{bonded}} + E_{\text{non.bonded}}$$

$$V(R) = E_{\text{bonded}} + E_{\text{non.bonded}}$$

Sum of **bonded** and **non-bonded**  
atom-type and position based terms

$$V(R) = E_{\text{bonded}} + E_{\text{non.bonded}}$$

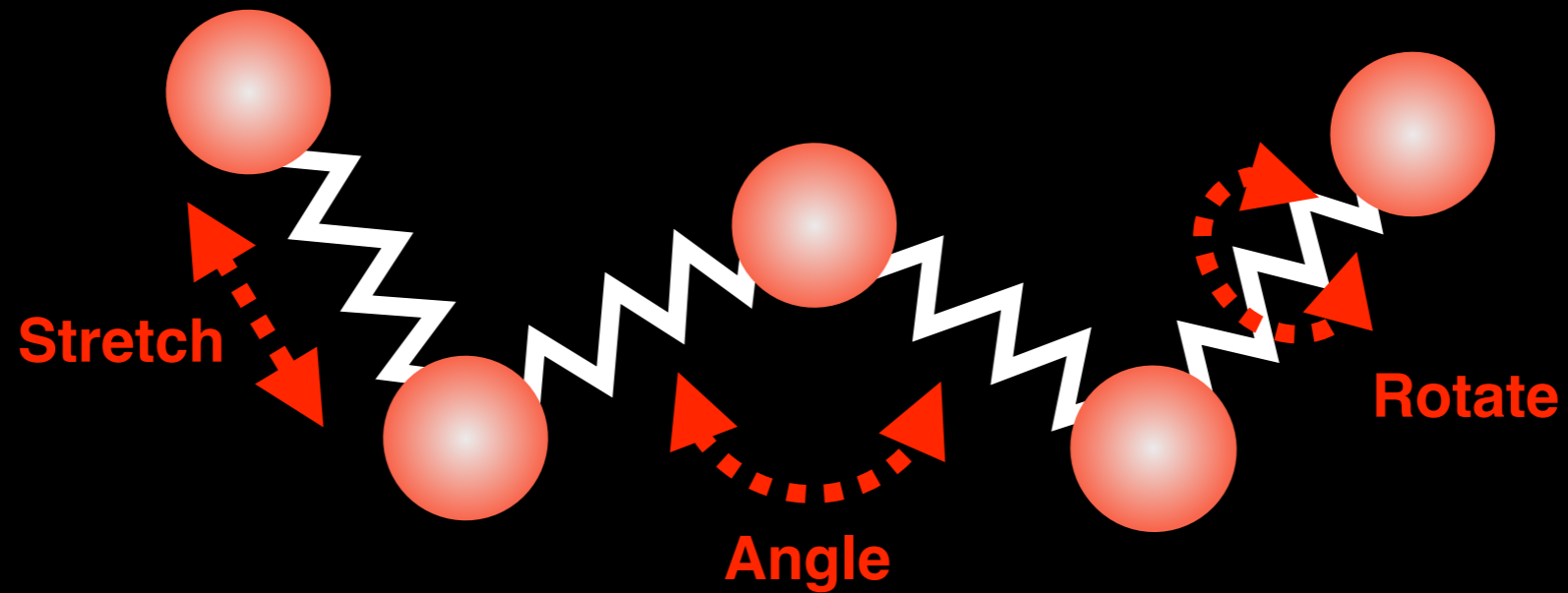
$E_{\text{bonded}}$  is itself a sum of three terms:



$$V(R) = E_{\text{bonded}} + E_{\text{non.bonded}}$$

$E_{\text{bonded}}$  is itself a sum of three terms:

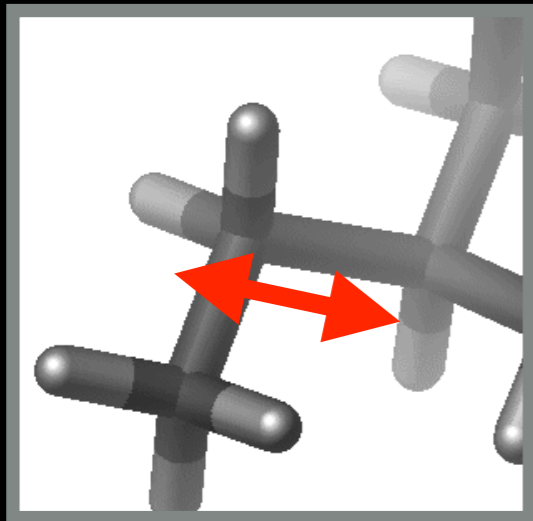
$$E_{\text{bond.stretch}} + E_{\text{bond.angle}} + E_{\text{bond.rotate}}$$



$$V(R) = E_{\text{bonded}} + E_{\text{non.bonded}}$$

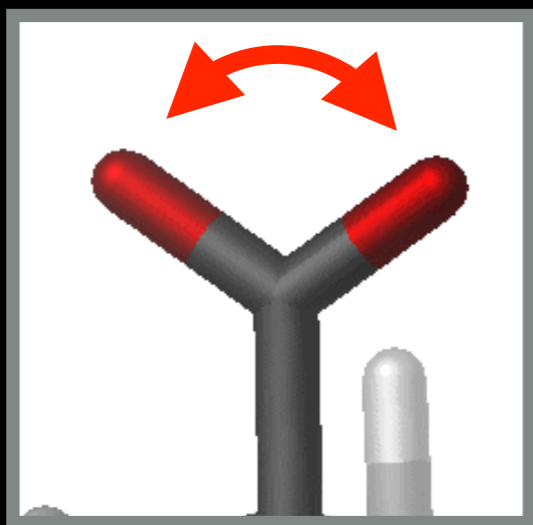
$E_{\text{bonded}}$  is itself a sum of three terms:

$$E_{\text{bond.stretch}} + E_{\text{bond.angle}} + E_{\text{bond.rotate}}$$



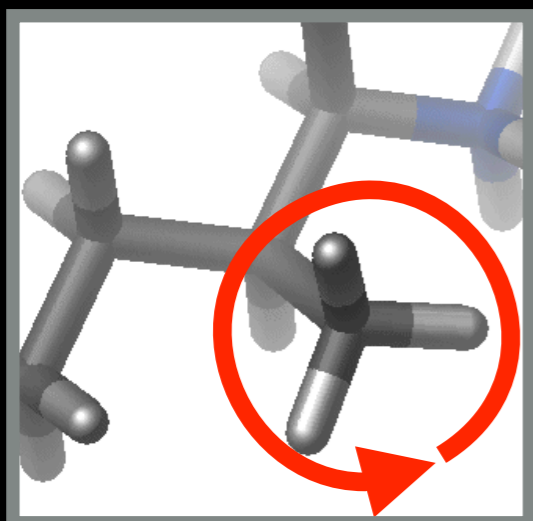
Bond Stretch

$E_{bond.stretch}$



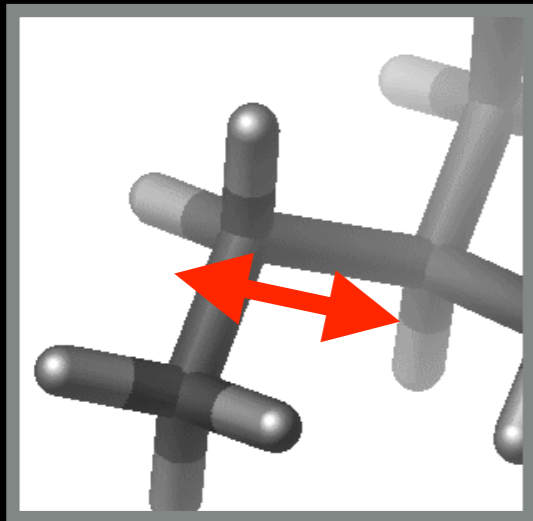
Bond Angle

$E_{bond.angle}$



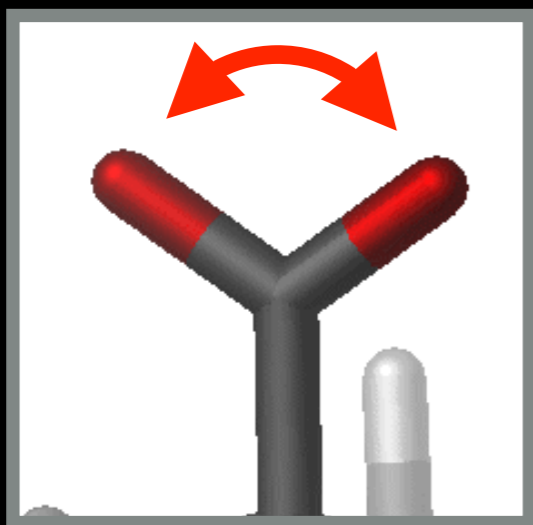
Bond Rotate

$E_{bond.rotate}$



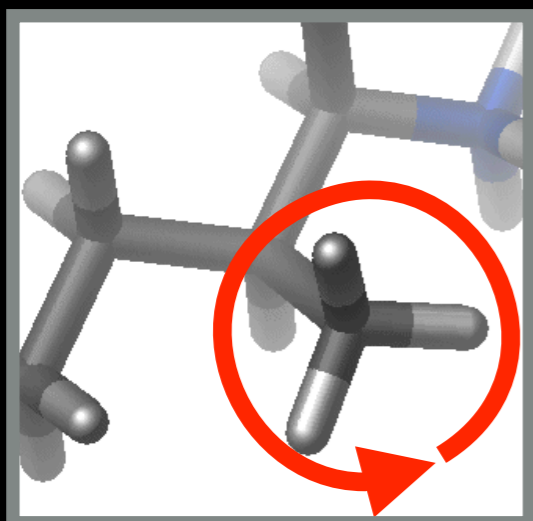
## Bond Stretch

$$\sum_{\text{bonds}} K_i^{bs} (b_i - b_o)$$



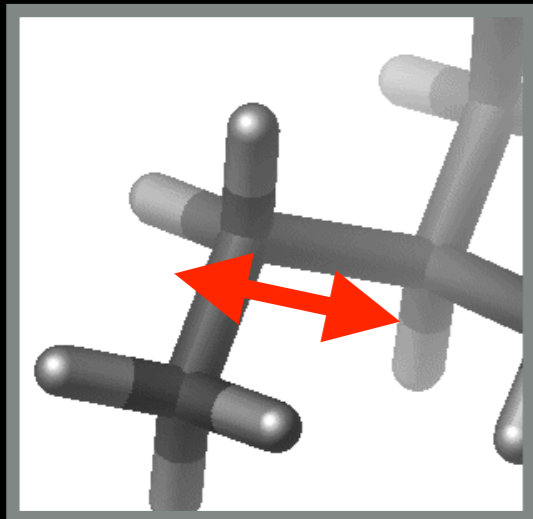
## Bond Angle

$$\sum_{\text{angles}} K_i^{ba} (\theta_i - \theta_o)$$



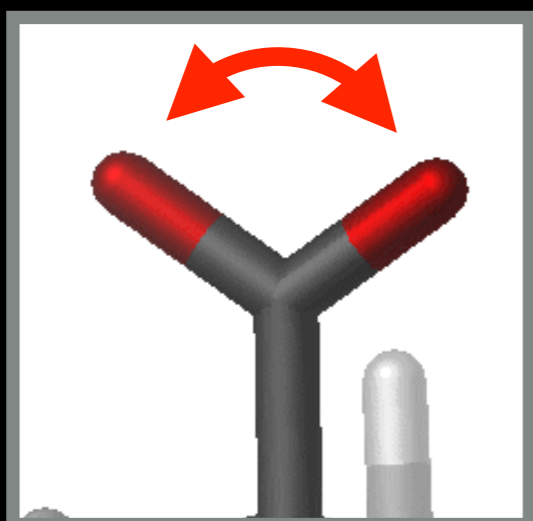
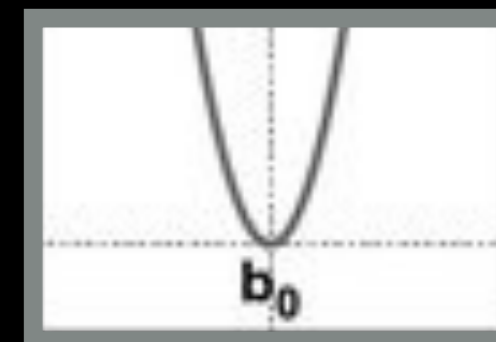
## Bond Rotate

$$\sum_{\text{dihedrals}} K_i^{br} [1 - \cos(n_i \phi_i - \phi_o)]$$



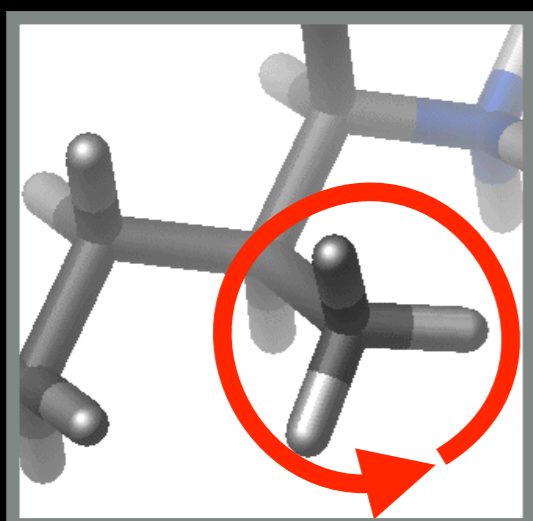
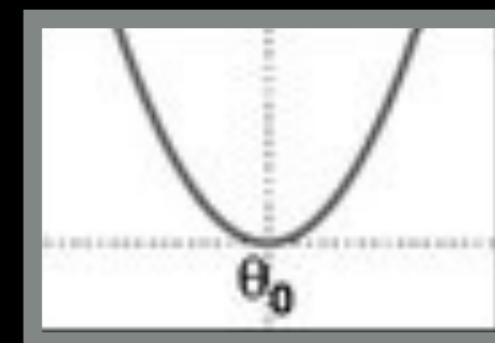
## Bond Stretch

$$\sum_{\text{bonds}} K_i^{bs} (b_i - b_o)$$



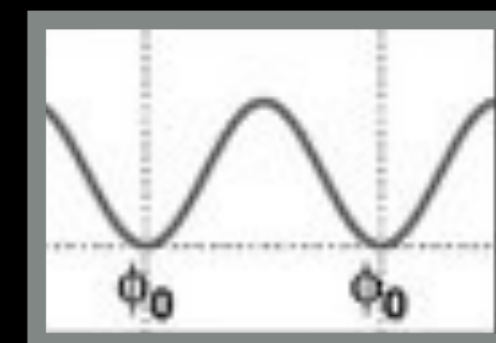
## Bond Angle

$$\sum_{\text{angles}} K_i^{ba} (\theta_i - \theta_o)$$



## Bond Rotate

$$\sum_{\text{dihedrals}} K_i^{br} [1 - \cos(n_i \phi_i - \phi_o)]$$



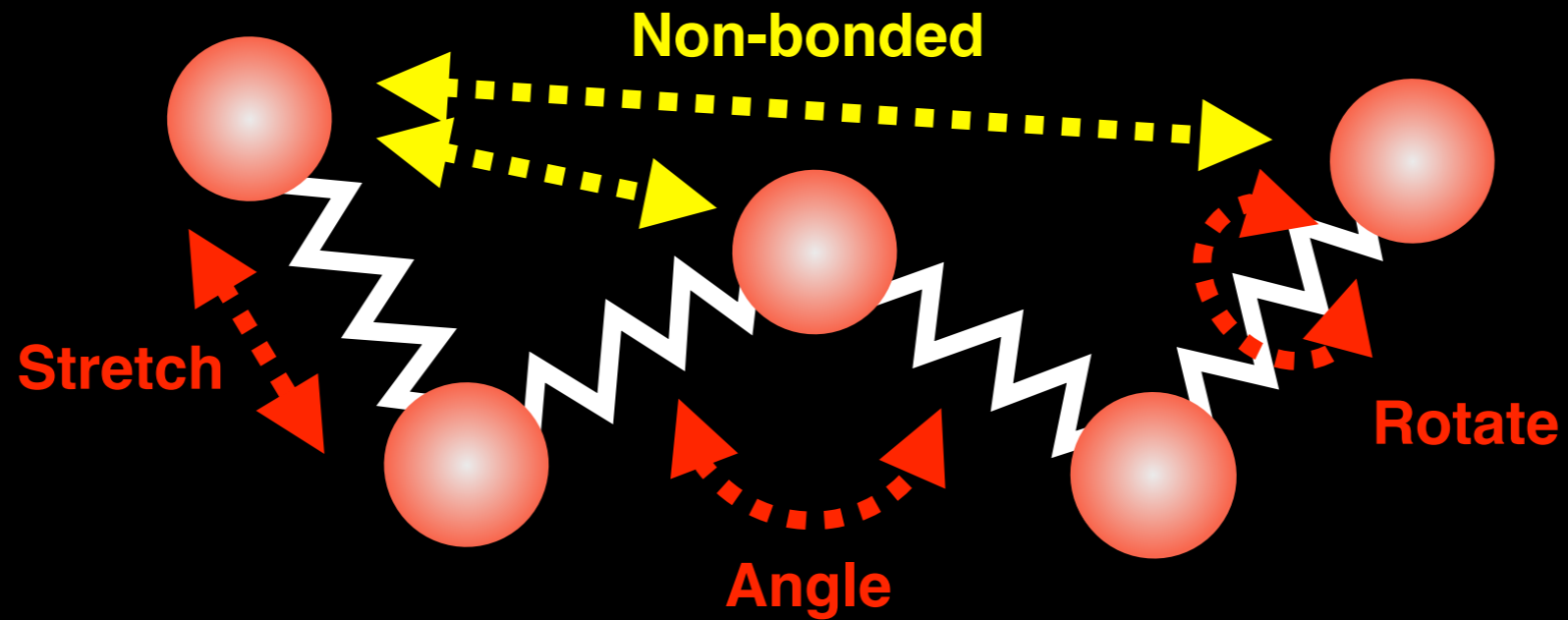
$$V(R) = E_{\text{bonded}} + E_{\text{non.bonded}}$$

$E_{\text{non.bonded}}$  is a sum of two terms:

$$V(R) = E_{\text{bonded}} + E_{\text{non.bonded}}$$

$E_{\text{non.bonded}}$  is a sum of two terms:

$$E_{\text{van.der.Waals}} + E_{\text{electrostatic}}$$

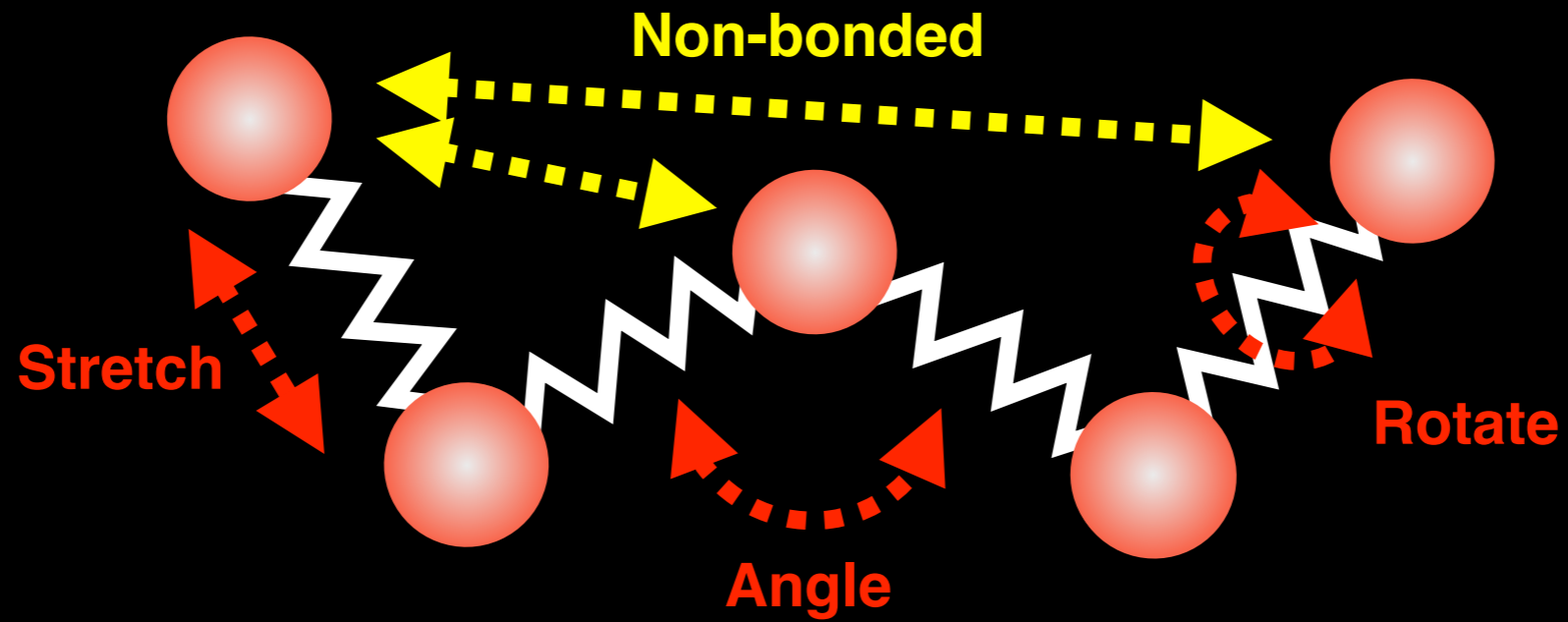


$$V(R) = E_{\text{bonded}} + E_{\text{non.bonded}}$$

$E_{\text{non.bonded}}$  is a sum of two terms:

$$E_{\text{van.der.Waals}} + E_{\text{electrostatic}}$$

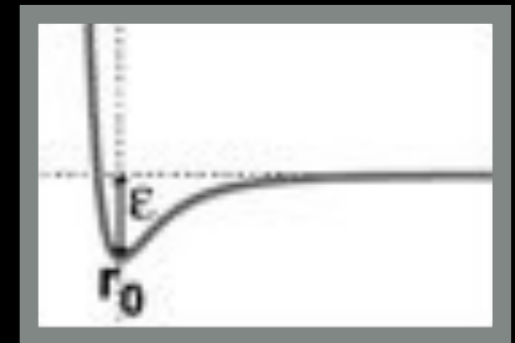




$$E_{electrostatic} = \sum_{pairs.i.j} \frac{q_i q_j}{\epsilon r_{ij}}$$



$$E_{van.der.Waals} = \sum_{pairs.i.j} \left[ \epsilon_{ij} \left( \frac{r_{o.ij}}{r_{ij}} \right)^{12} - 2\epsilon_{ij} \left( \frac{r_{o.ij}}{r_{ij}} \right)^6 \right]$$



# Total potential energy

The potential energy can be given as a sum of terms for: Bond stretching, Bond angles, Bond rotations, van der Waals and Electrostatic interactions between atom pairs

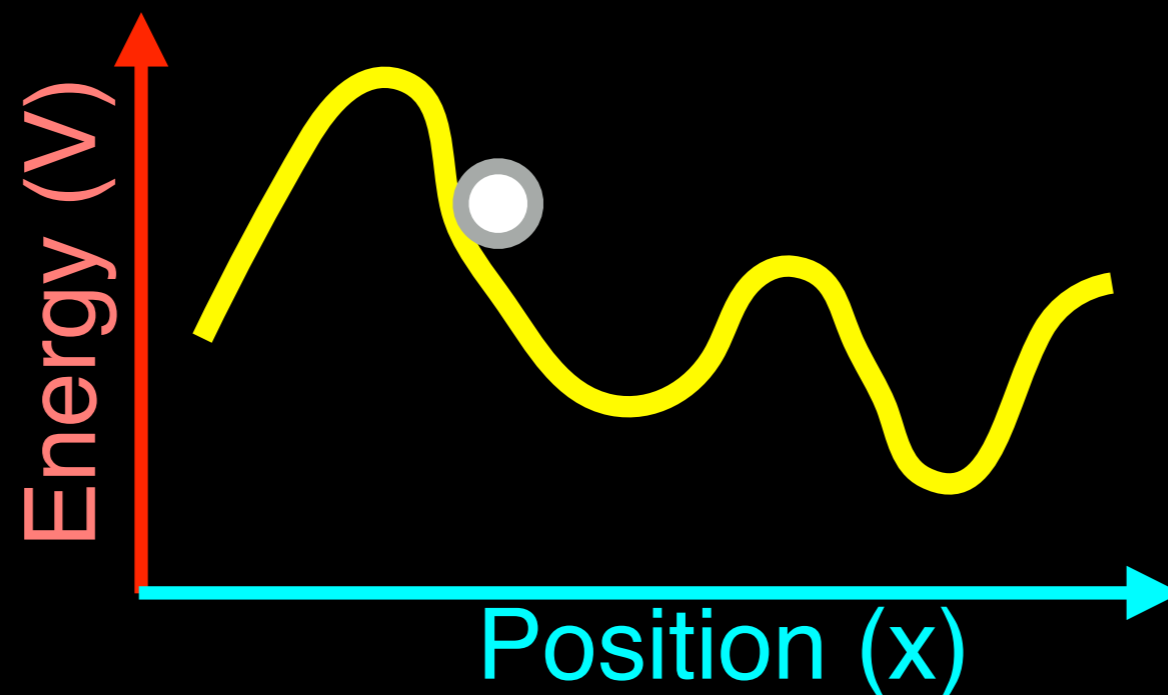
$$V(R) = E_{\text{bond.stretch}} + E_{\text{bond.angle}} + E_{\text{bond.rotate}} + E_{\text{van.der.Waals}} + E_{\text{electrostatic}}$$

$E_{\text{bonded}}$

$E_{\text{non.bonded}}$

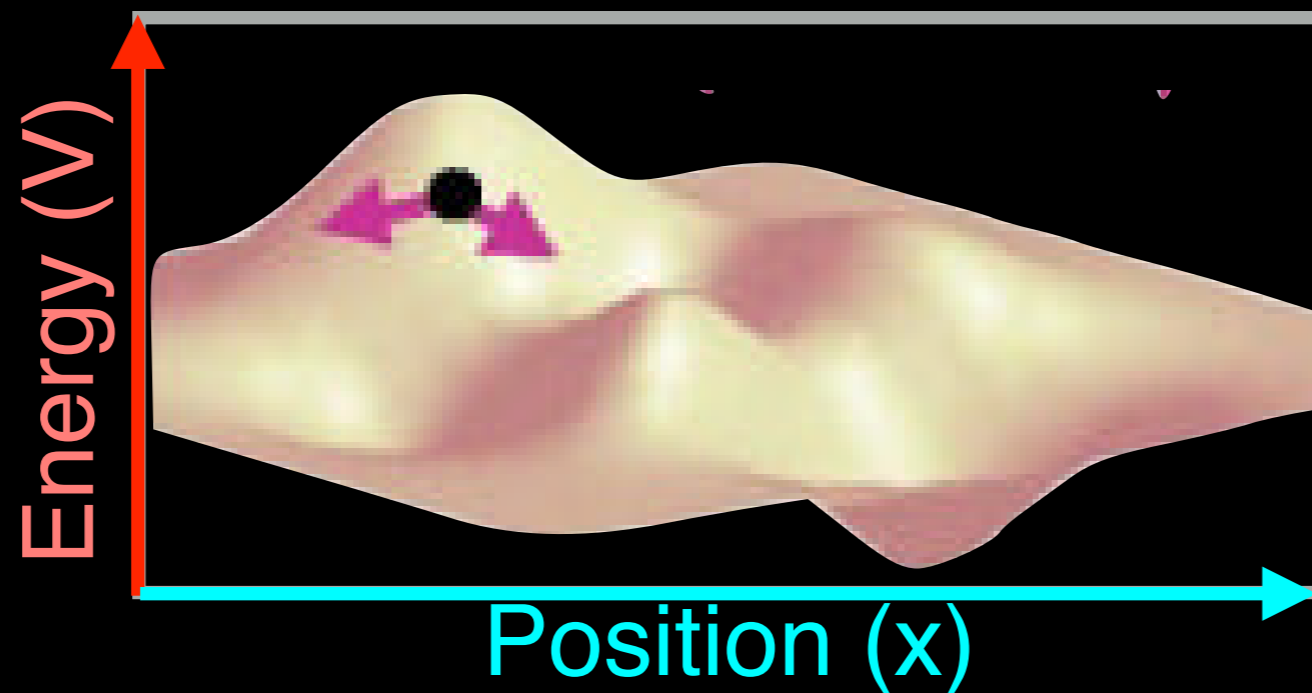
# Potential energy surface

Now we can calculate the **potential energy surface** that fully describes the energy of a molecular system as a function of its geometry



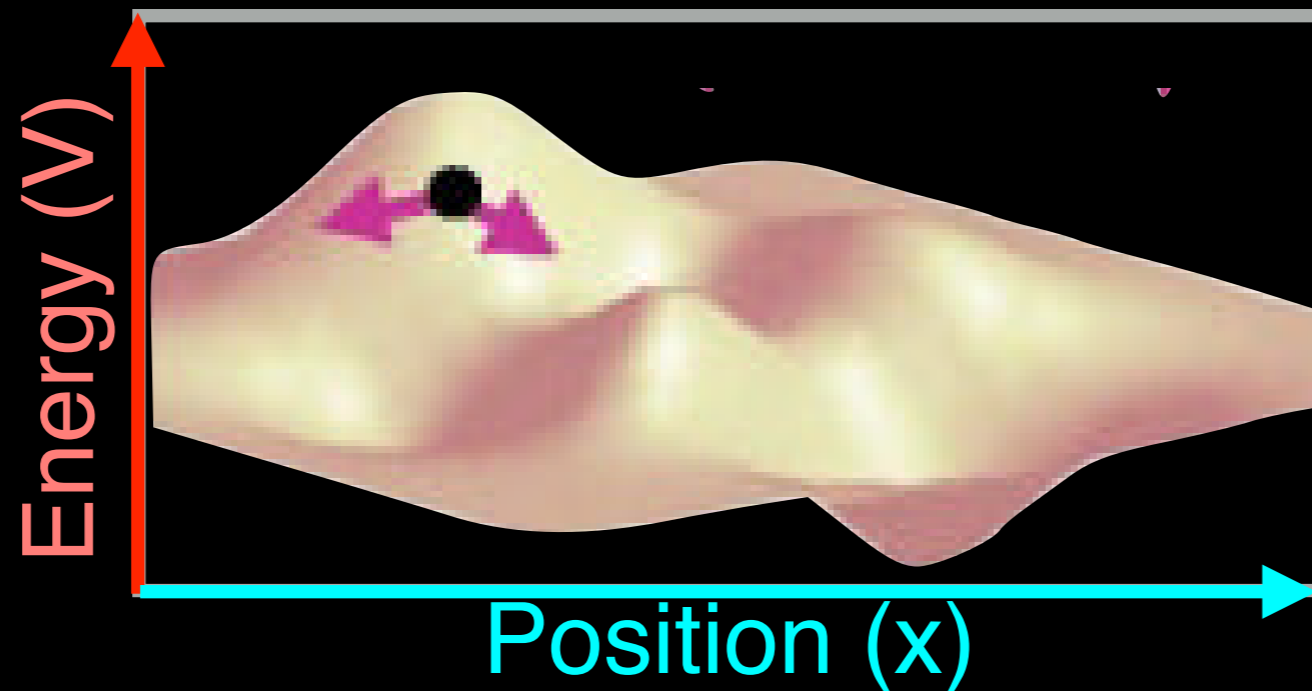
# Potential energy surface

Now we can calculate the **potential energy surface** that fully describes the energy of a molecular system as a function of its geometry



# Key concept:

Now we can calculate the **potential energy surface** that fully describes the energy of a molecular system as a function of its geometry



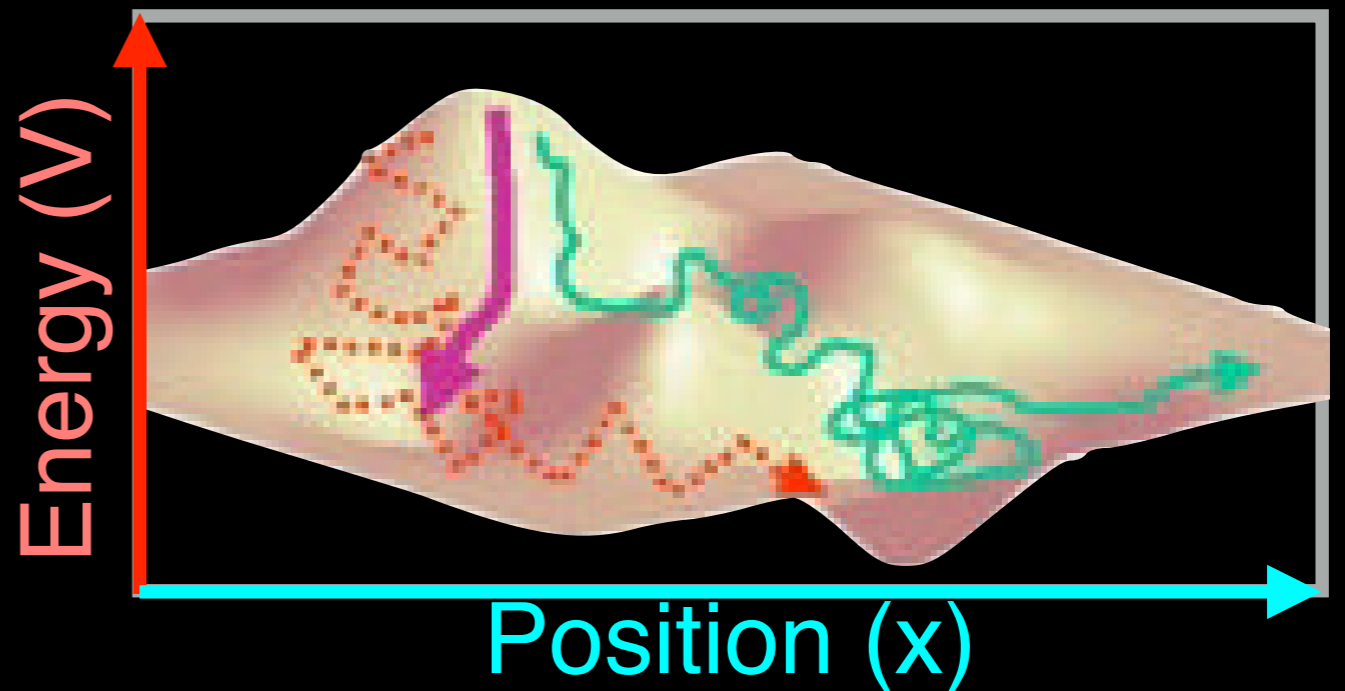
- The **forces** are the gradients of the energy

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

# 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:

$$\exp(-\Delta V/dx)$$



# 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

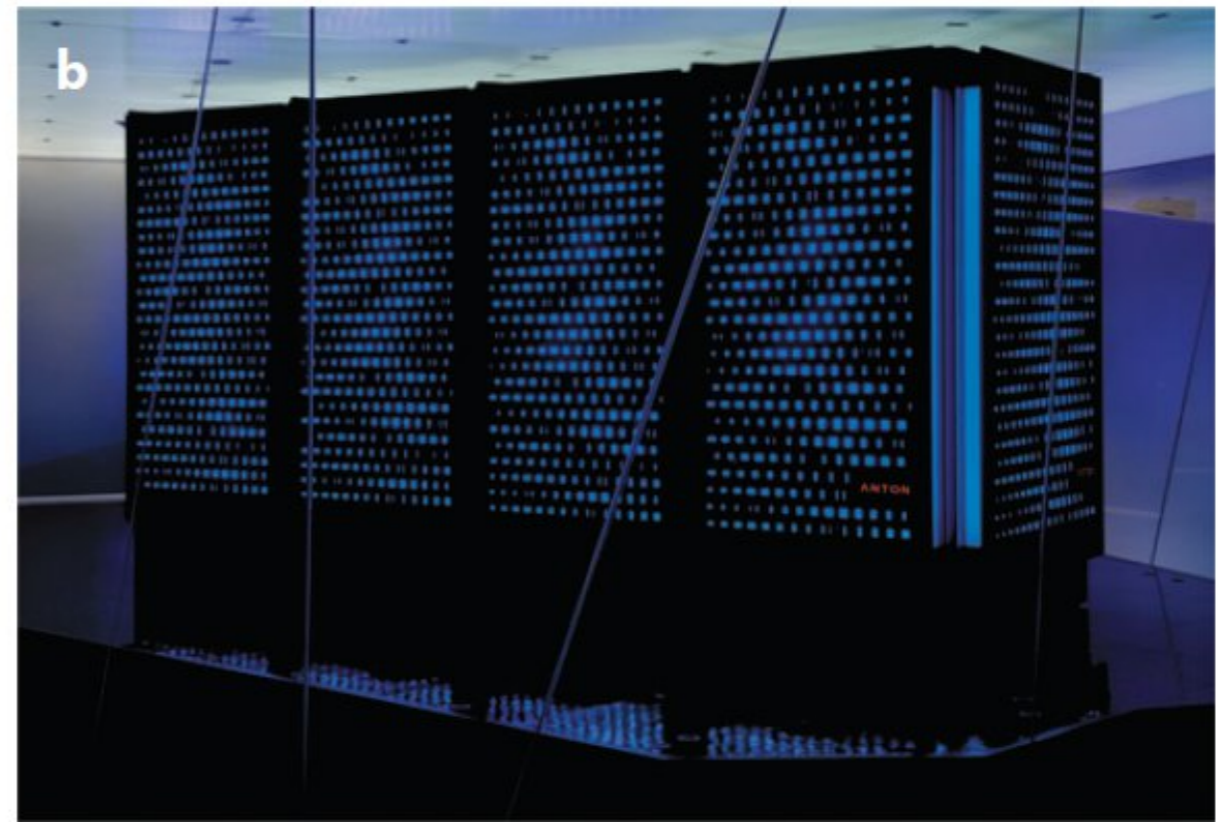
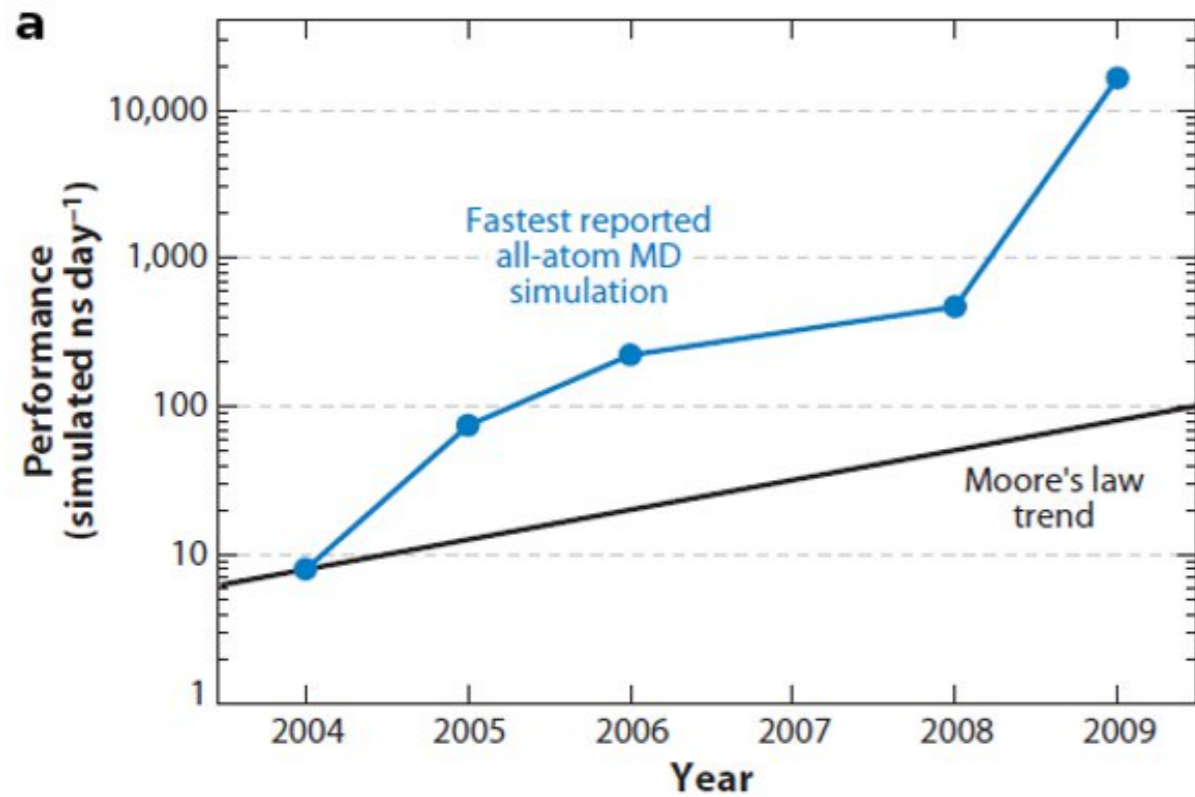
DATE	COST	SPEED	MEMORY	SIZE
1967	\$40M	0.1 MHz	1 MB	HALL
2013	\$4,000	1 GHz	10 GB	LAPTOP
CHANGE	10,000	10,000	10,000	10,000

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

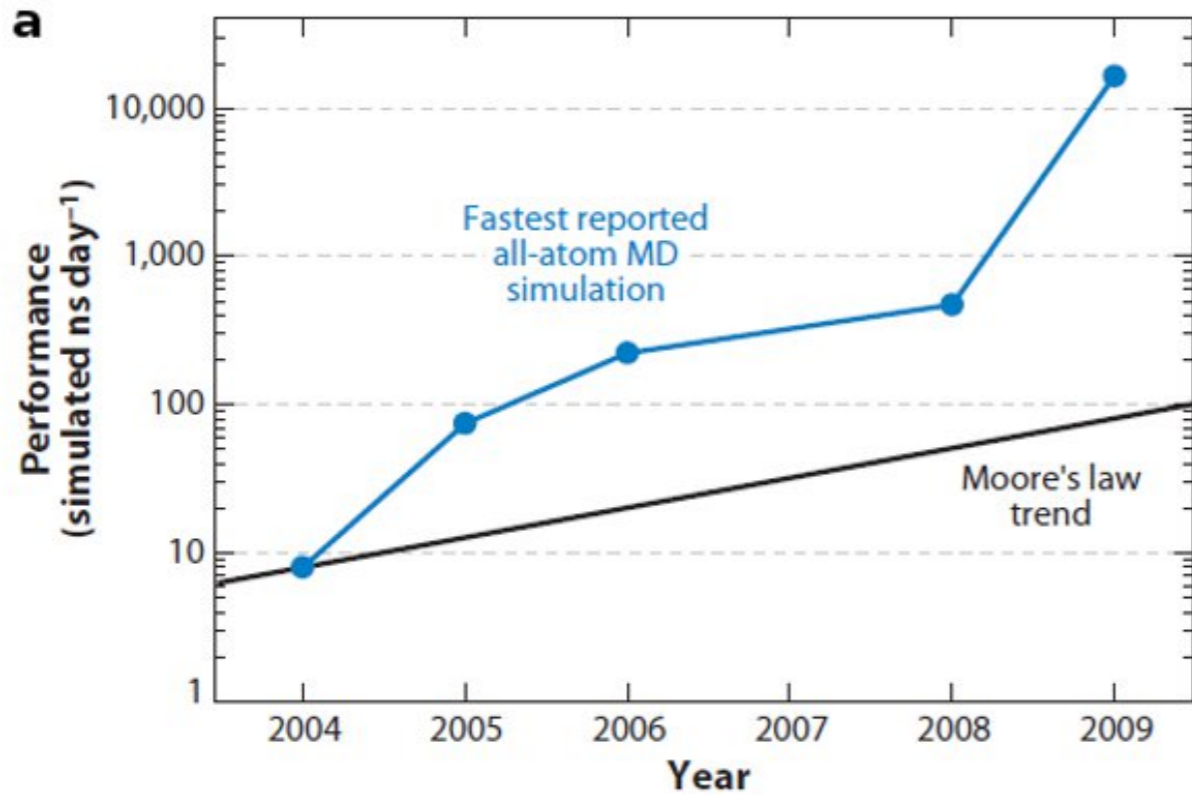




# SIDE-NOTE: GPUS AND ANTON SUPERCOMPUTER



# SIDE-NOTE: GPUS AND ANTON SUPERCOMPUTER



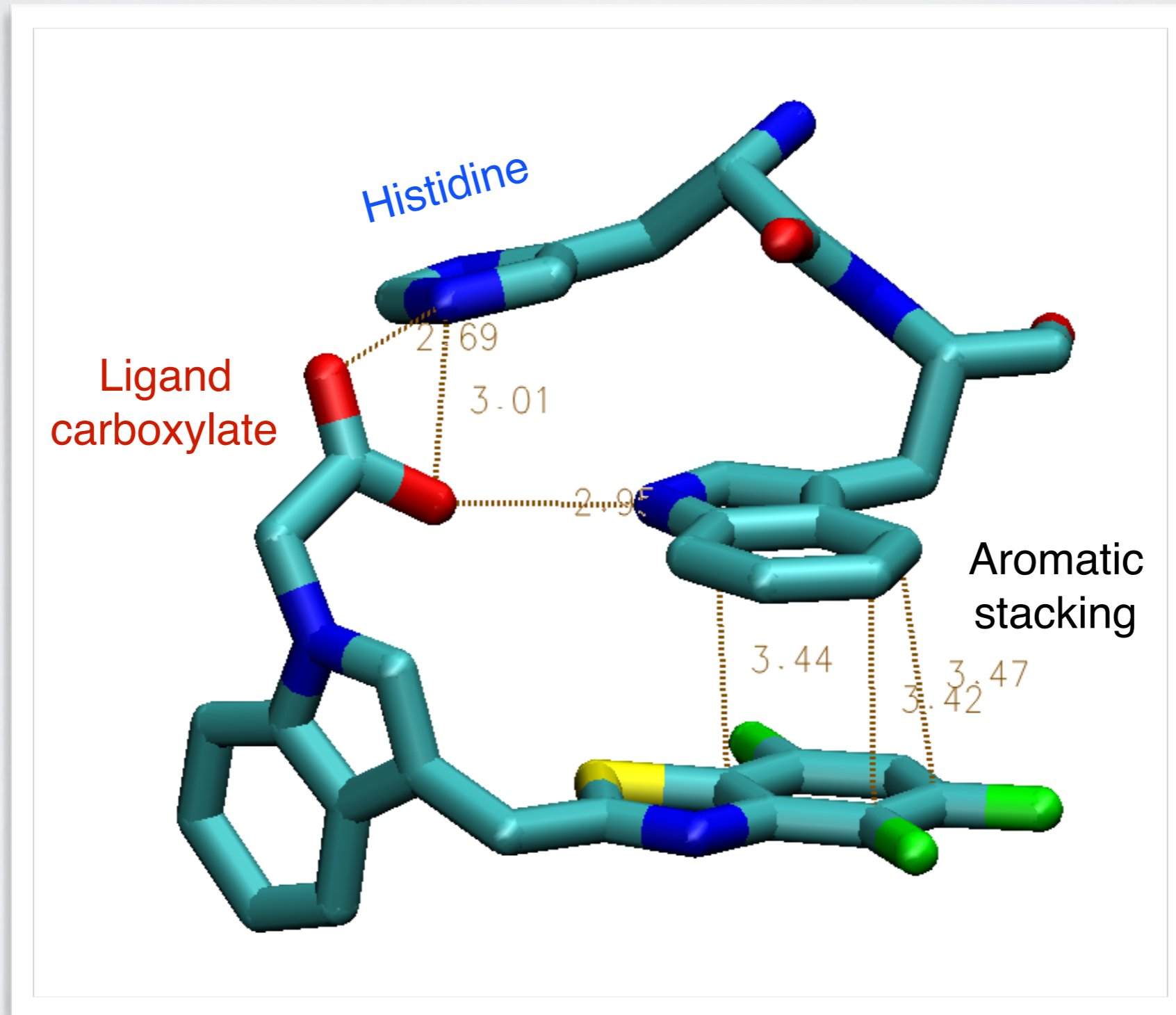
POTENTIAL FUNCTIONS DESCRIBE A SYSTEMS  
**ENERGY** AS A FUNCTION OF ITS **STRUCTURE**

Two main approaches:

(1). **Physics-Based**

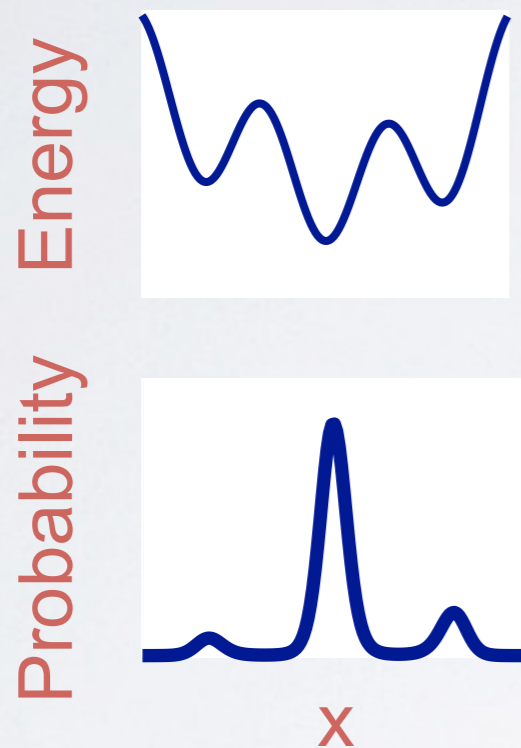
(2). **Knowledge-Based**

# KNOWLEDGE-BASED DOCKING POTENTIALS



# ENERGY DETERMINES **PROBABILITY** (STABILITY)

Basic idea: Use probability as a proxy for energy



Boltzmann:

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

Inverse Boltzmann:

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

Example: ligand **carboxylate O** to protein **histidine N**

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

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

# 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)

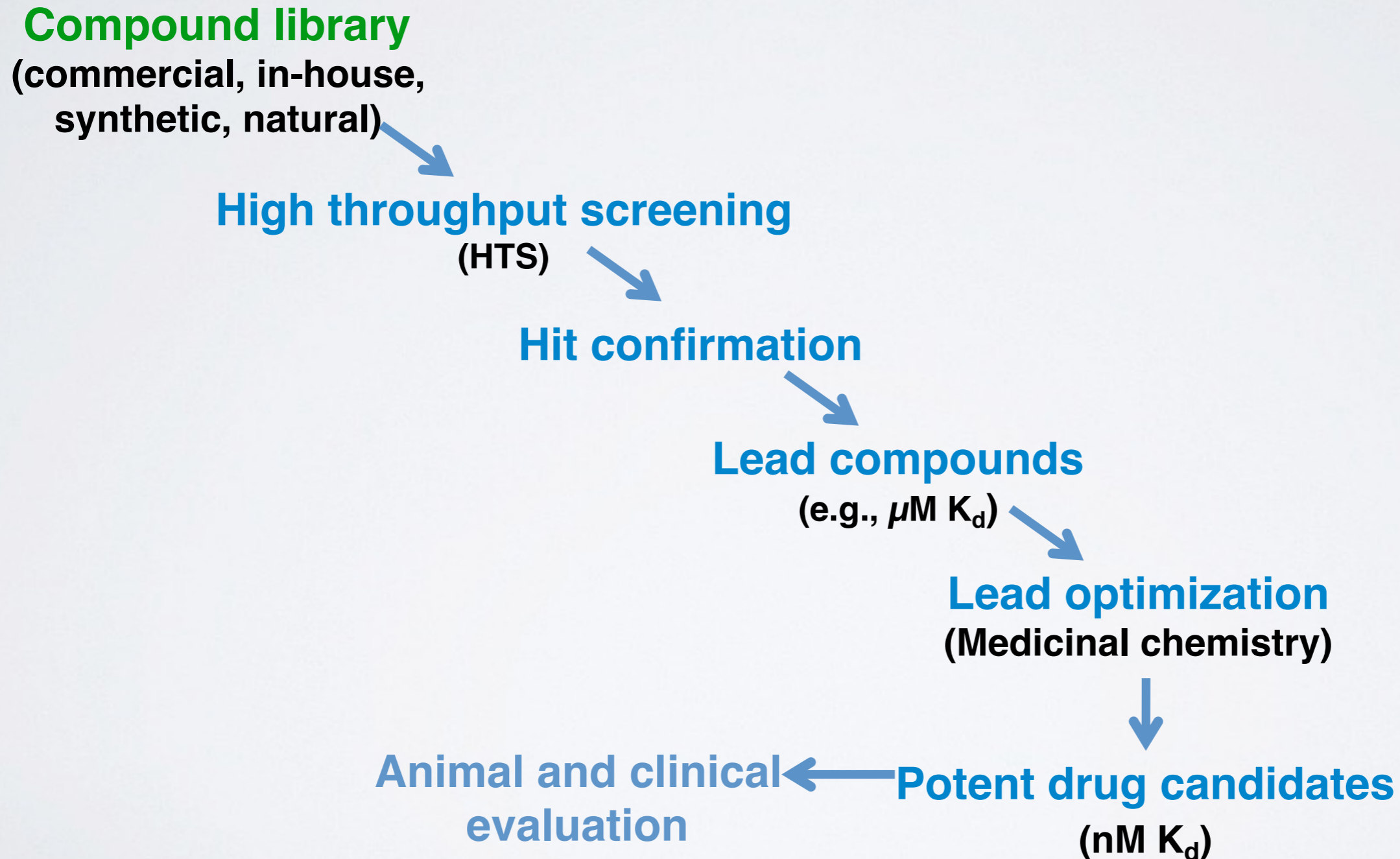
# Computer Aided Drug Discovery

# Next Up:

- **Overview of structural bioinformatics**
  - Motivations, goals and challenges
- **Fundamentals of protein structure**
  - Structure composition, form and forces
- **Representing, interpreting & modeling protein structure**
  - Visualizing and interpreting protein structures
  - Analyzing protein structures
  - Modeling energy as a function of structure
  - Drug discovery & Predicting functional dynamics



# THE TRADITIONAL EMPIRICAL PATH TO DRUG DISCOVERY



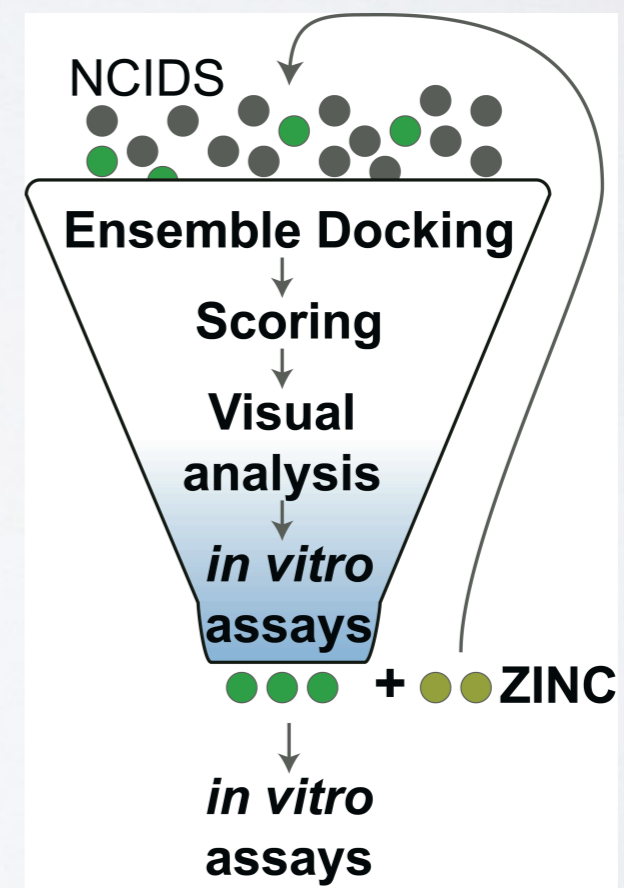
# COMPUTER-AIDED LIGAND DESIGN

Aims to reduce number of compounds synthesized and assayed

Lower costs

Reduce chemical waste

Facilitate faster progress



Two main approaches:

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

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

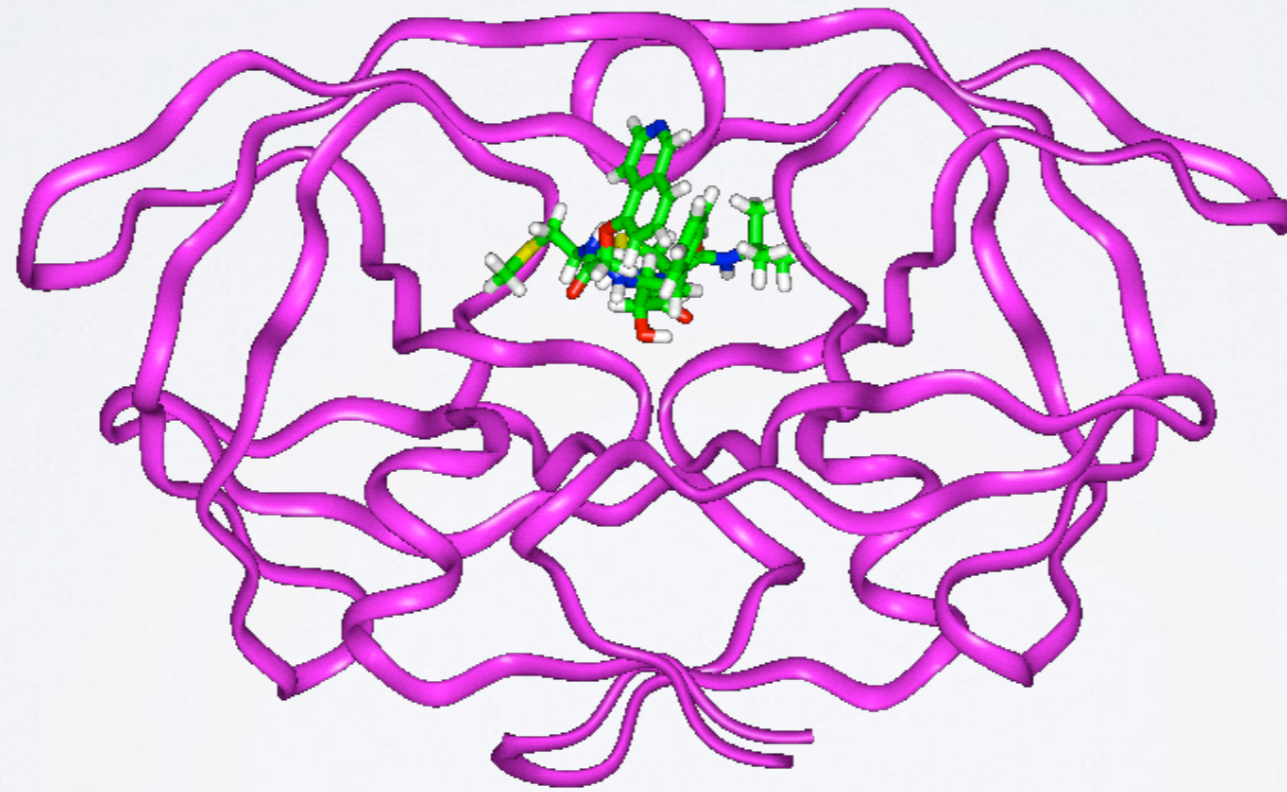
Two main approaches:

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

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

# SCENARIO I: RECEPTOR-BASED DRUG DISCOVERY

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



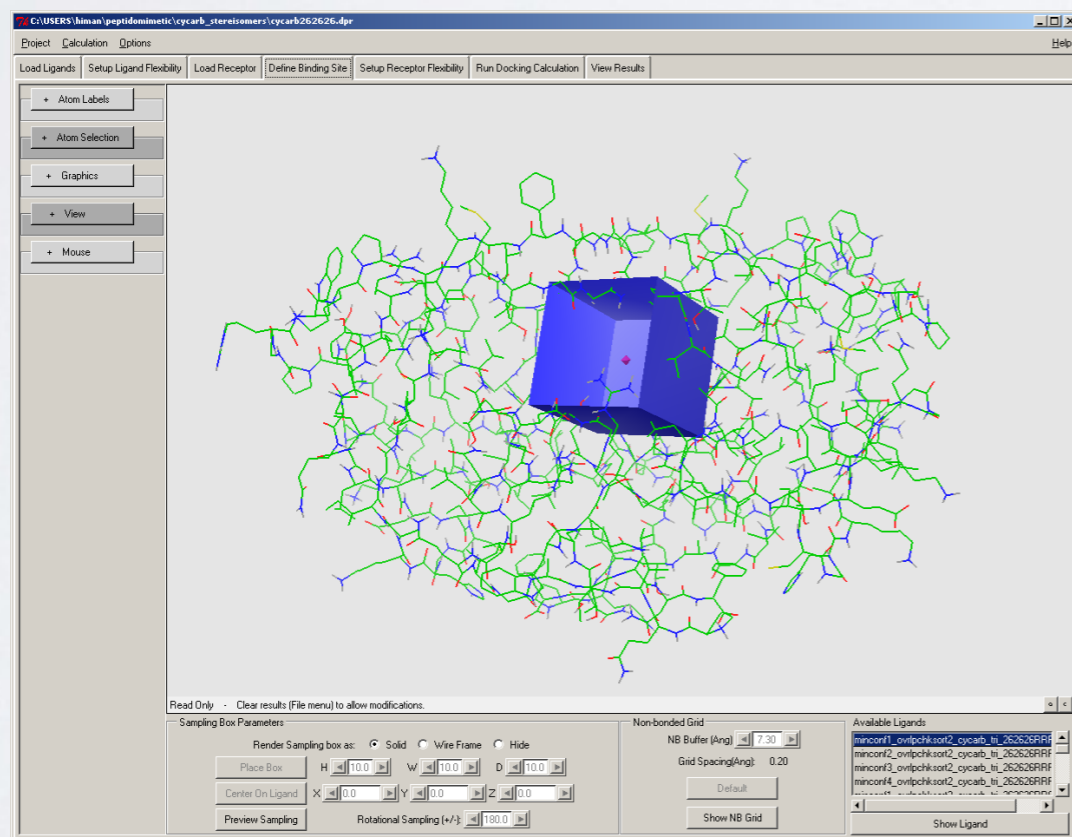
HIV Protease/KNI-272 complex

# PROTEIN-LIGAND DOCKING

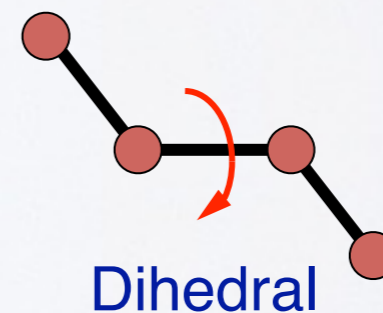
## Structure-Based Ligand Design

### Docking software

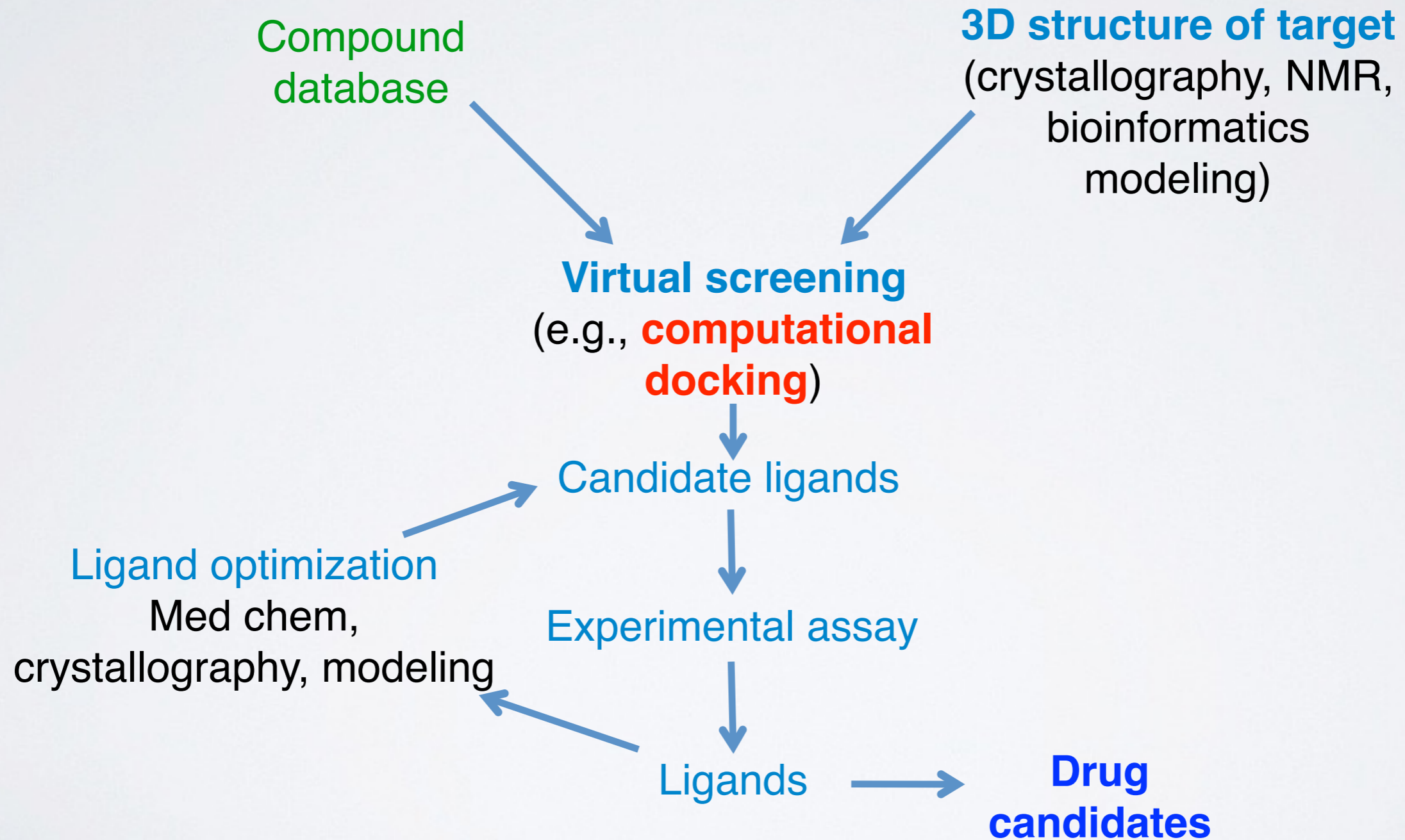
Search for structure of lowest energy



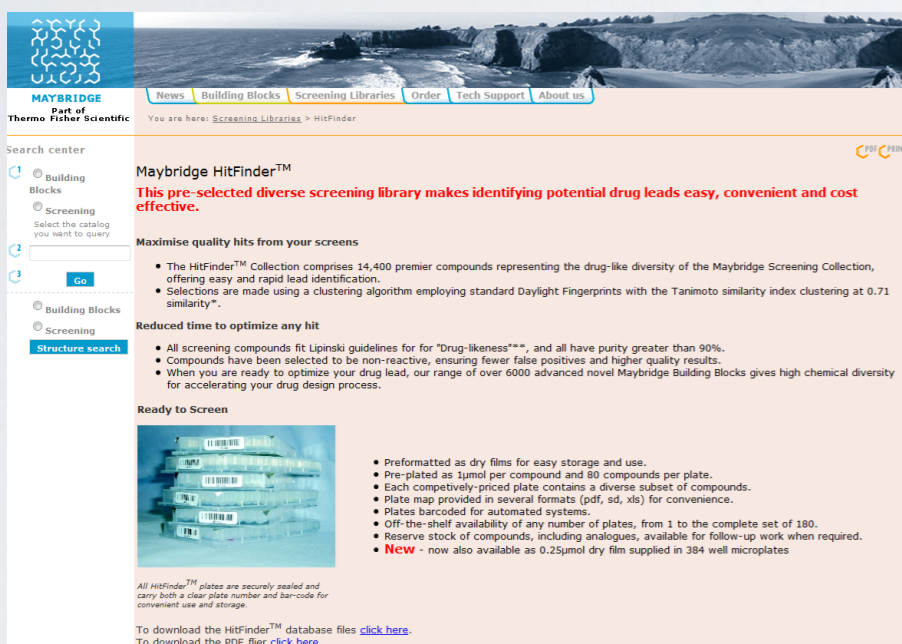
Potential function  
Energy as function of structure



# STRUCTURE-BASED VIRTUAL SCREENING



# COMPOUND LIBRARIES



The screenshot shows the Maybridge HitFinder website, which is part of Thermo Fisher Scientific. The page features a search center with options for Building Blocks, Screening, and Structure search. The main content area highlights the Maybridge HitFinder™ collection, describing it as a pre-selected diverse screening library. It lists key features such as 14,400 premier compounds, ease of use, and high quality results. A 'Ready to Screen' section includes a list of benefits like preformatted dry films, 1µmol per compound, and barcoding. The footer contains copyright information for Galapagos NV and links to download database files and PDFs.

Commercial  
(in-house pharma)



The screenshot displays the NIH Molecular Libraries Small Molecule Repository website, operated by BioFocus, a Galapagos Company. The page is titled 'A Roadmap Initiative' and features a 'Welcome' message. It describes the repository's role in collecting samples for high-throughput biological screening and distributing them to the NIH Molecular Libraries Probe Production Centers Network. A 'Registered Users Login' section is visible on the left. The footer includes copyright information for Galapagos NV and a note that BioFocus, a Galapagos company, operates MLSMR in South San Francisco.

Government (NIH)



The screenshot shows the Pittsburgh Molecular Libraries Screening Center (PMLSC) website, part of the University of Pittsburgh. The page features a navigation menu with links to Home, History, Personnel, Screening Technology, Compound Libraries, Instrumentation/Platforms, HTS Guidelines, Approved PMLSC Assay Protocols, PMLSC Probe Reports, Chemistry, Data Analysis/Informatics, Educational Activities, Publications, Links, and Contacts. A search bar is present. The main content area includes a 'Welcome' message and a description of the center's mission to assist scientists and the National Institutes of Health in interrogating small molecule libraries using optical-based High Throughput and High Content assays. The footer contains links to Health Sciences @ Pitt, UPMC, HSLS, School of Medicine, Health Sciences Calendar, and Our News & Events, along with copyright information for the University of Pittsburgh.

Academia



# COMMON SIMPLIFICATIONS USED IN PHYSICS-BASED DOCKING

Quantum effects approximated classically

Protein often held rigid

Configurational entropy neglected

Influence of water treated crudely

Do it Yourself!

# Hand-on time!

[https://bioboot.github.io/bgggn213\\_W19/lectures/#12](https://bioboot.github.io/bgggn213_W19/lectures/#12)

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

Two main approaches:

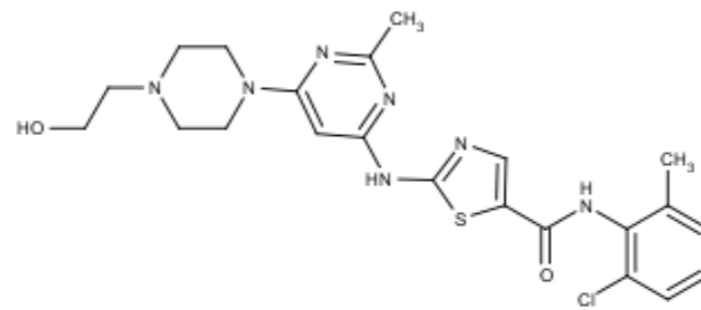
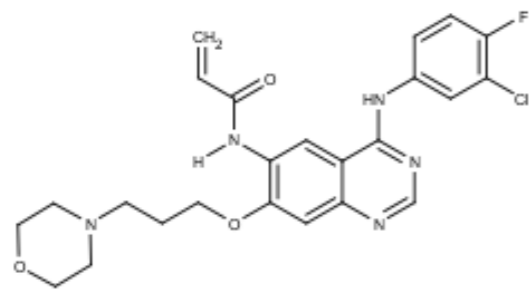
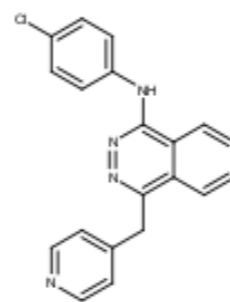
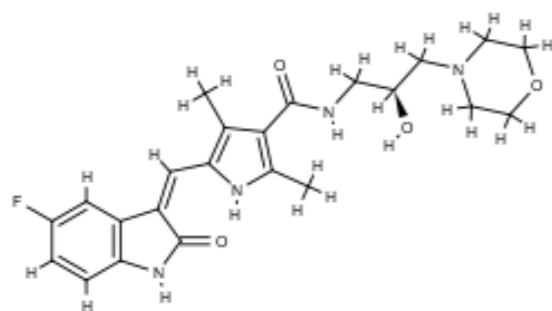
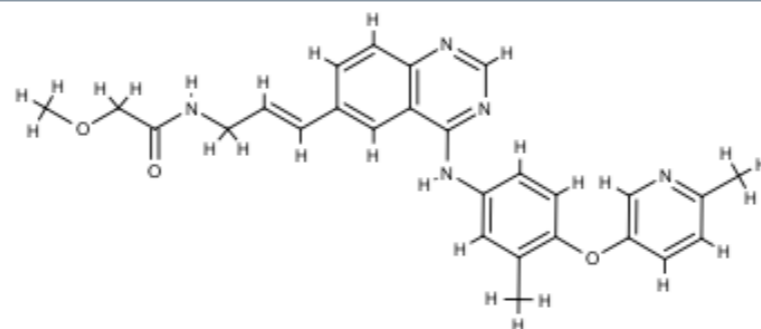
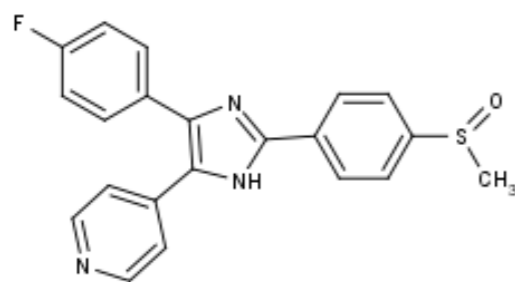
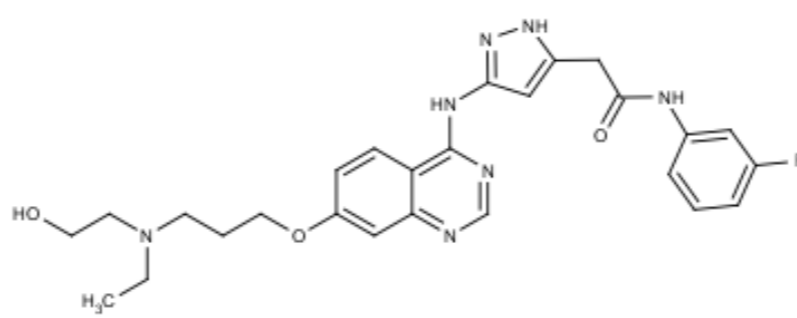
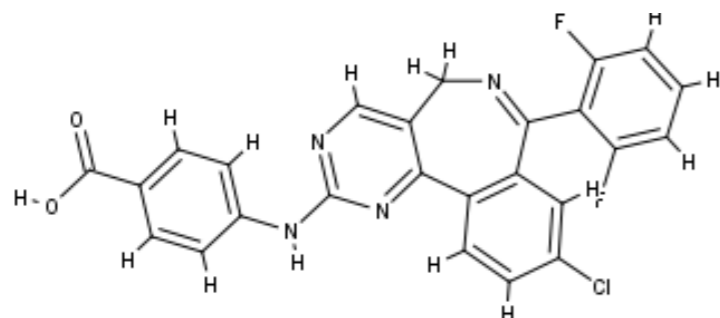
(1). **Receptor/Target-Based**

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

# Scenario 2

## Structure of Targeted Protein Unknown: Ligand-Based Drug Discovery

e.g. MAP Kinase Inhibitors



Using knowledge of existing inhibitors to discover more

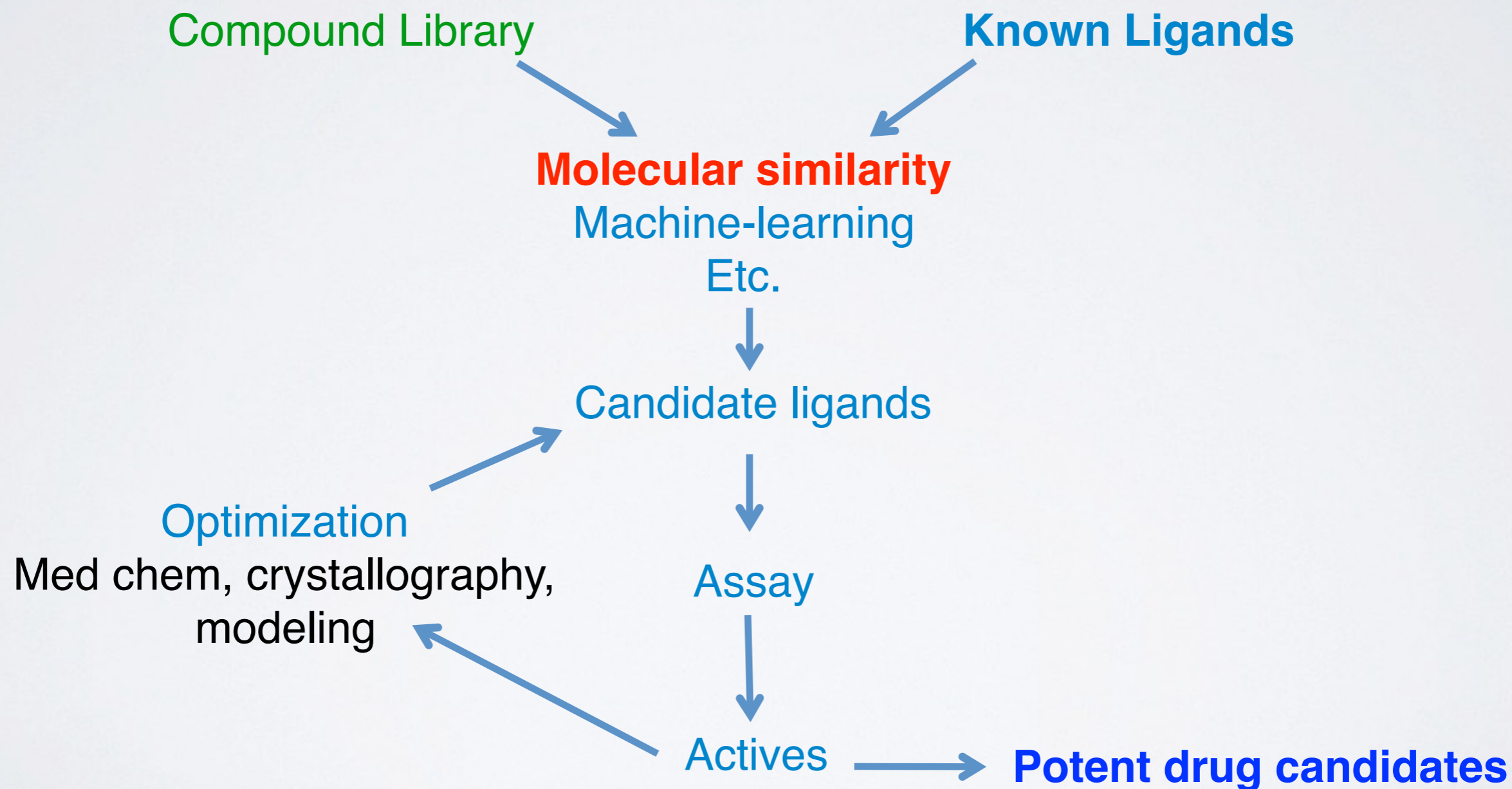
# Why Look for Another Ligand if You Already Have Some?

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

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

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

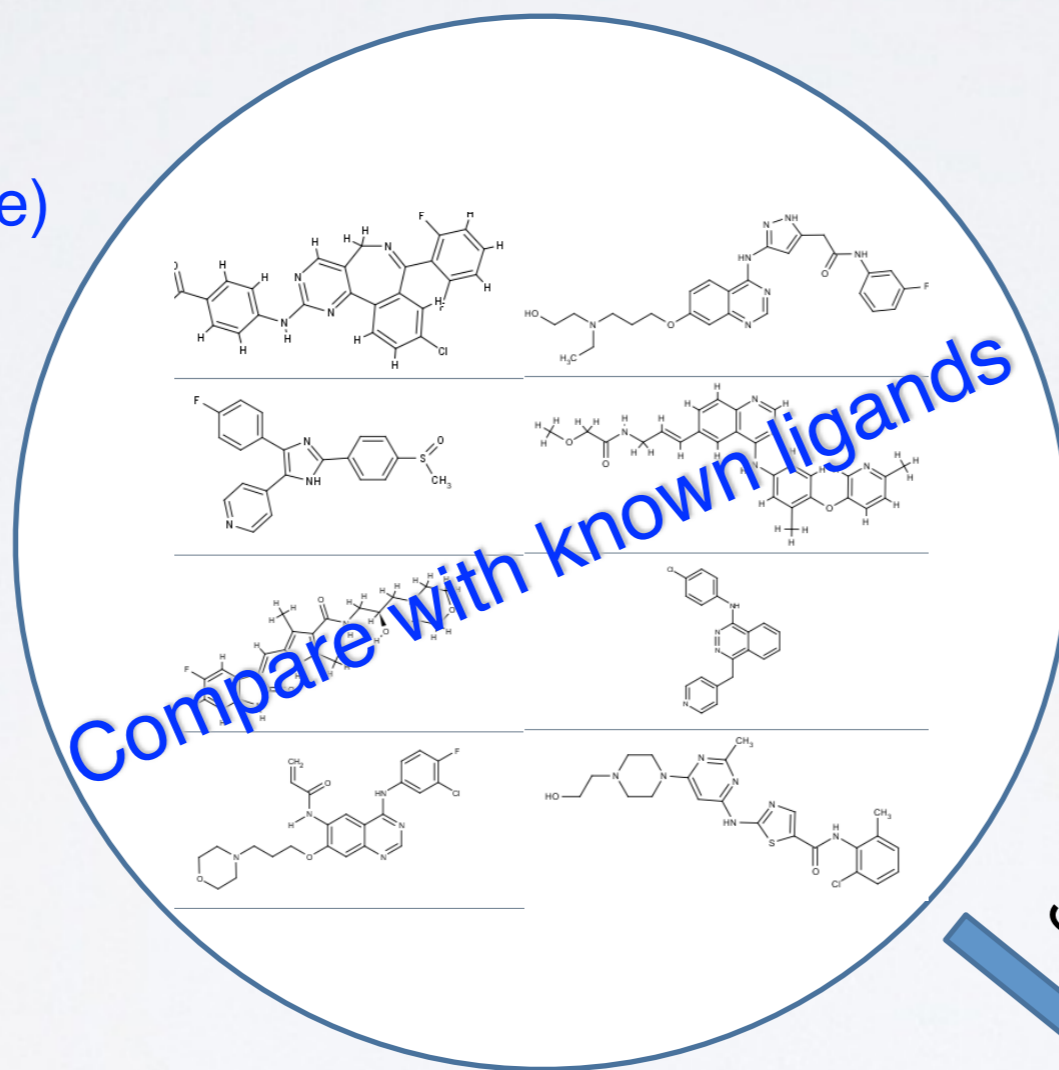
# LIGAND-BASED VIRTUAL SCREENING



# CHEMICAL SIMILARITY

## LIGAND-BASED DRUG-DISCOVERY

Compounds  
(available/synthesizable)



Different

Don't bother

Similar

Test experimentally

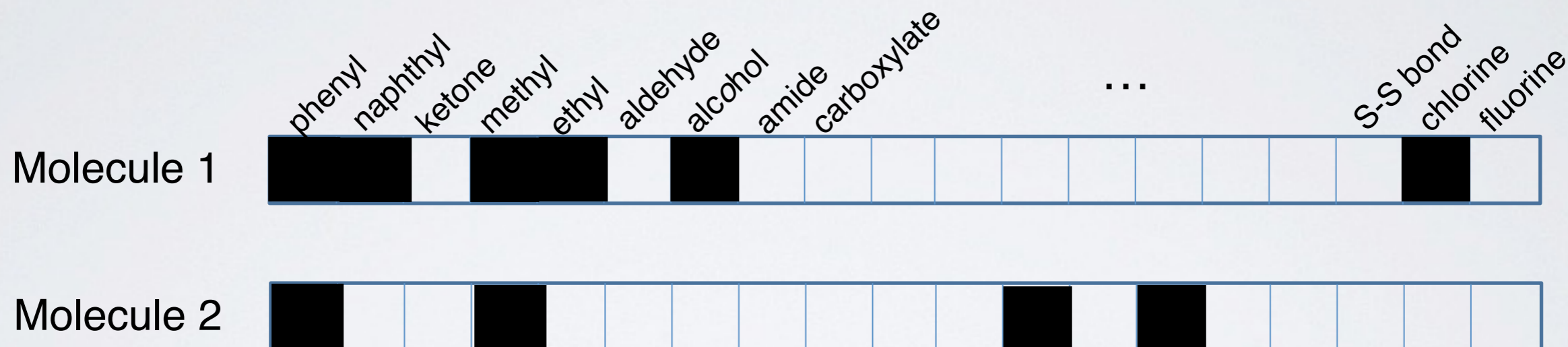
# CHEMICAL FINGERPRINTS

## BINARY STRUCTURE KEYS





# CHEMICAL SIMILARITY FROM FINGERPRINTS



Tanimoto Similarity  
(or Jaccard Index),  $T$

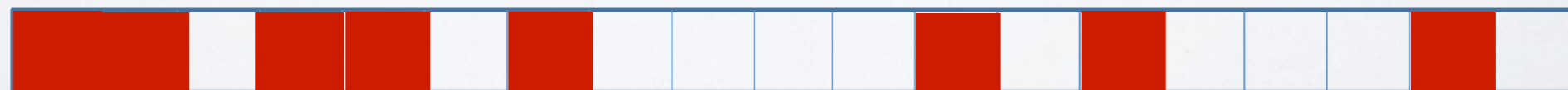
$$T \equiv \frac{N_I}{N_U} = 0.25$$

Intersection



$N_I=2$

Union

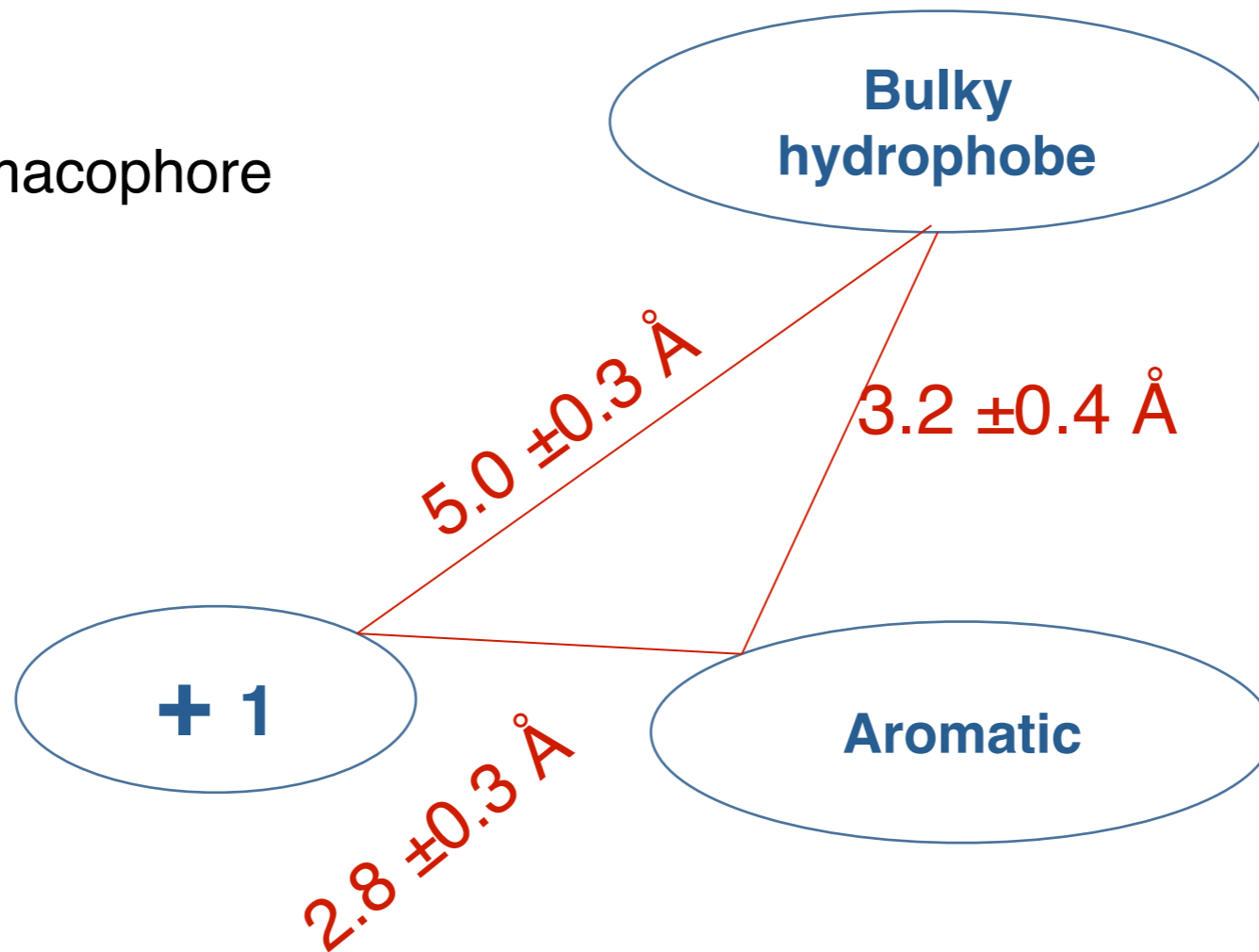


$N_U=8$

# Pharmacophore Models

Φάρμακο (drug) + Φορά (carry)

A 3-point pharmacophore



# Molecular Descriptors

More abstract than chemical fingerprints

## Physical descriptors

molecular weight

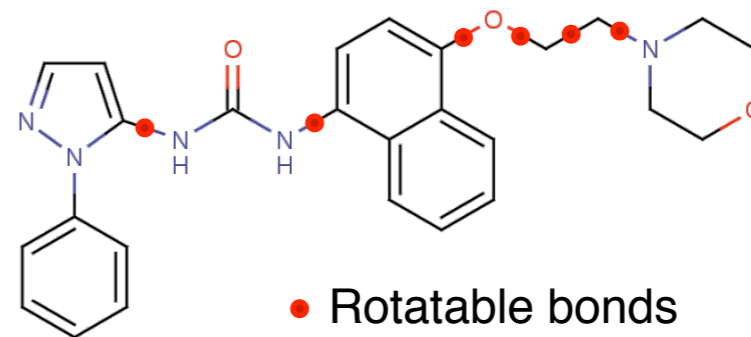
charge

dipole moment

number of H-bond donors/acceptors

number of rotatable bonds

hydrophobicity (log P and clogP)



## Topological

branching index

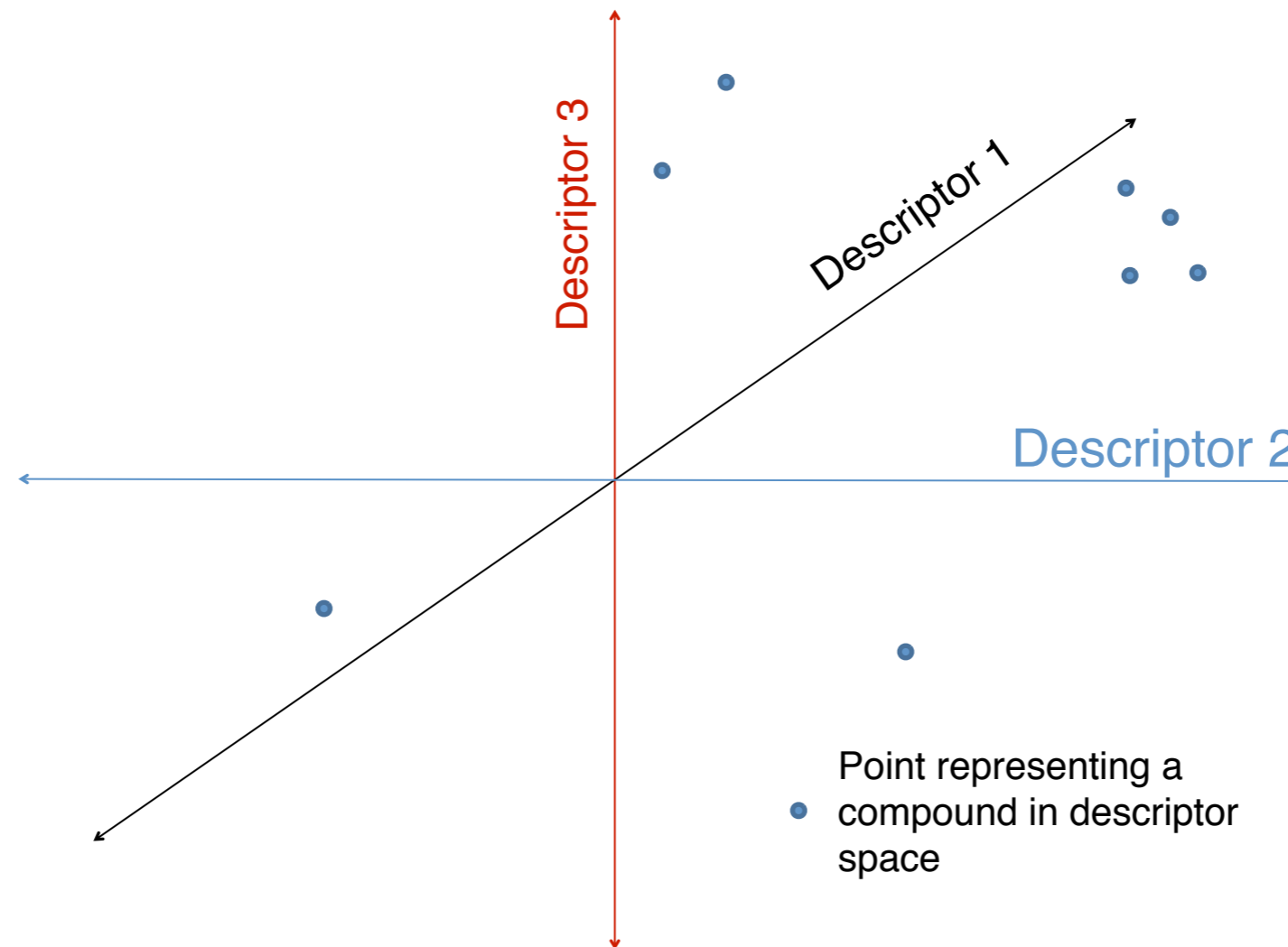
measures of linearity vs interconnectedness

Etc. etc.

# A High-Dimensional “Chemical Space”

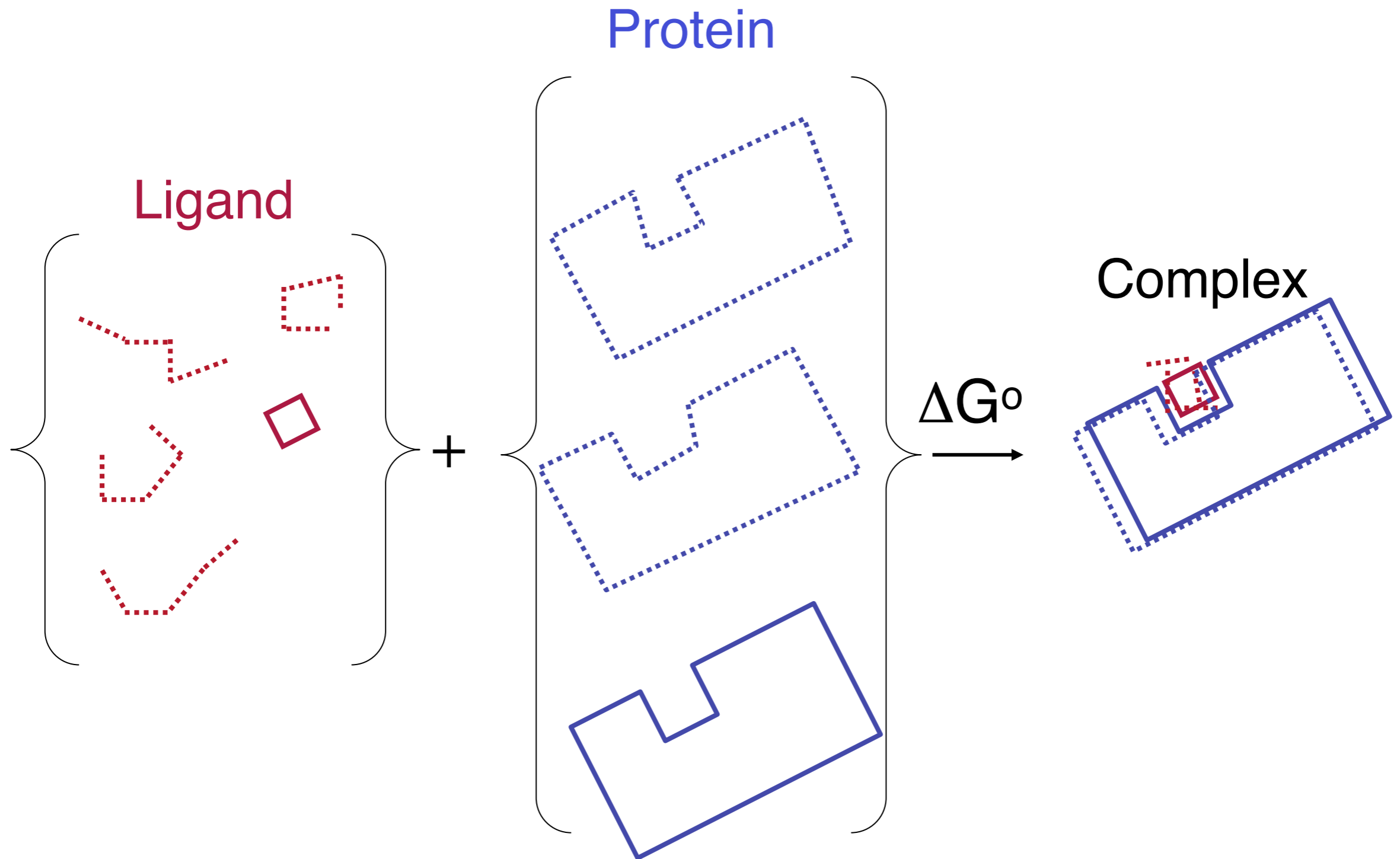
Each compound is a point in an n-dimensional space

Compounds with similar properties are near each other



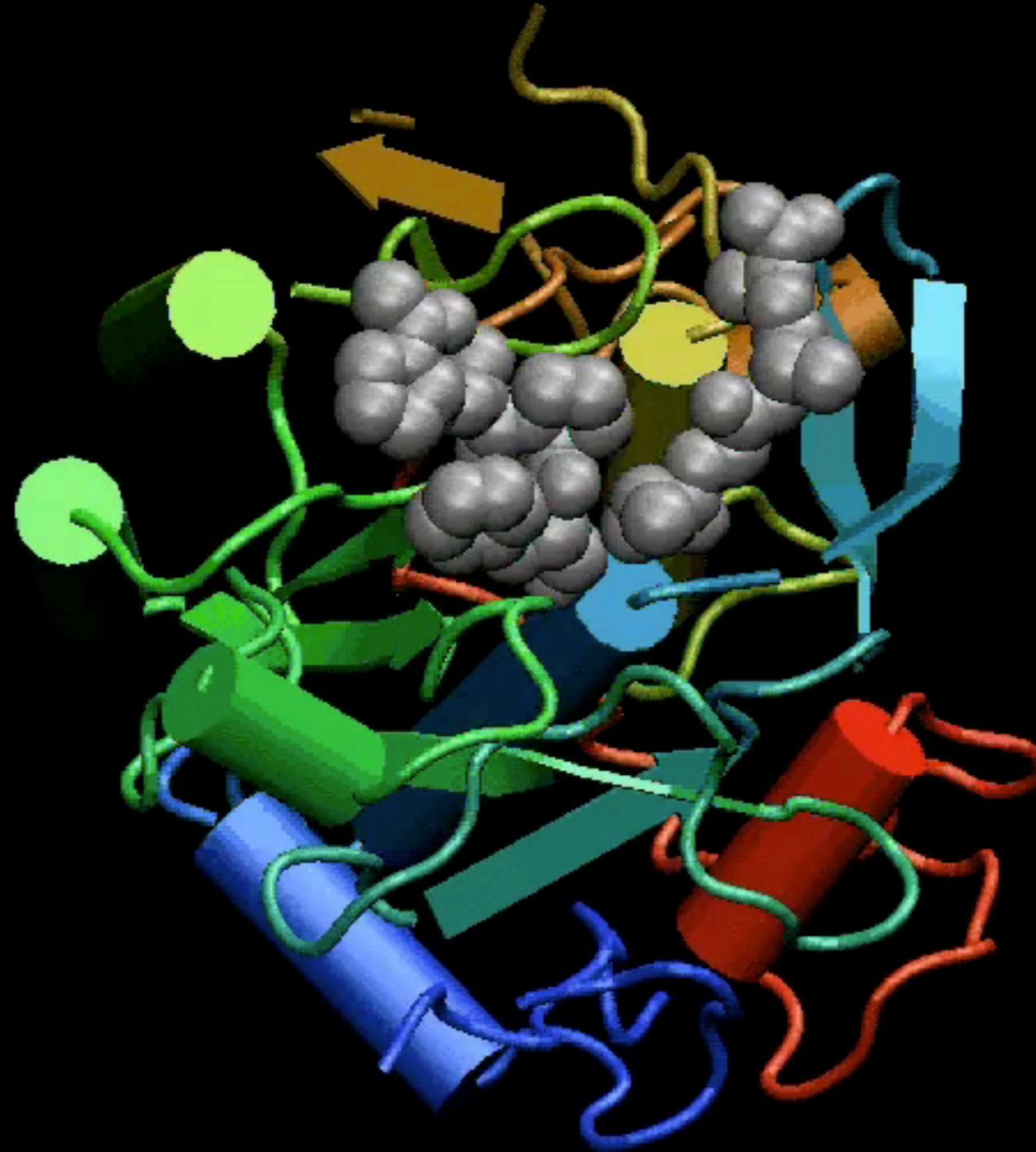
Apply **multivariate statistics** and **machine learning** for descriptor-selection. (e.g. partial least squares, PCA, support vector machines, random forest, deep learning etc.)

# Proteins and Ligand are Flexible



**NMA** (Normal Mode Analysis) is a bioinformatics method to predict the intrinsic dynamics of biomolecules

Do it Yourself!



[https://bioboot.github.io/bggn213\\_W19/lectures/#12](https://bioboot.github.io/bggn213_W19/lectures/#12)

# Reference Slides

Molecular Dynamics (MD) and Normal Mode Analysis  
(NMA) Background and Cautionary Notes

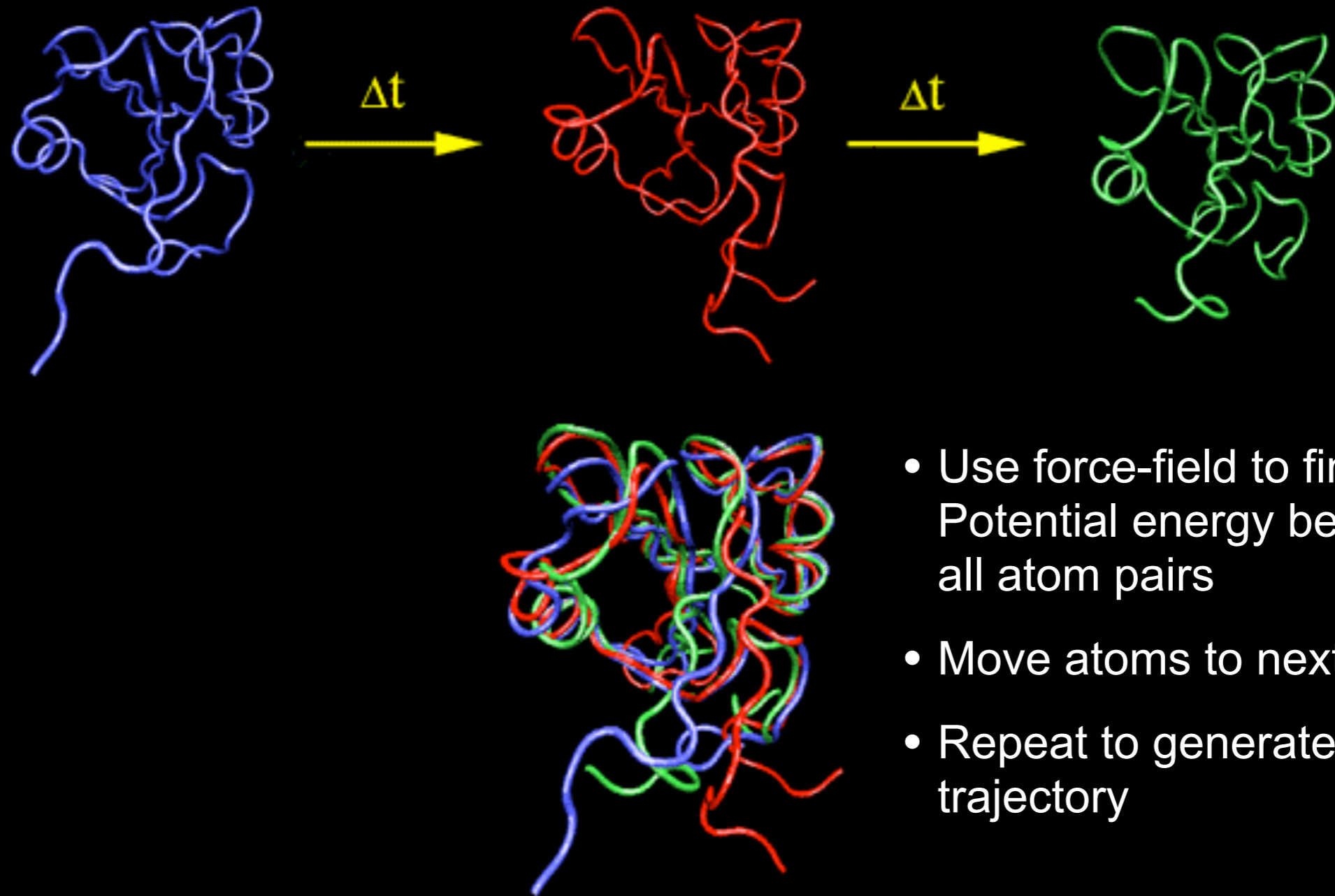
[ [Muddy Point Assessment](#) ]

# 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 ( $\sim 1$ fs) **time steps** ( $\Delta t$ )  
(for integrating equations of motion, see below)



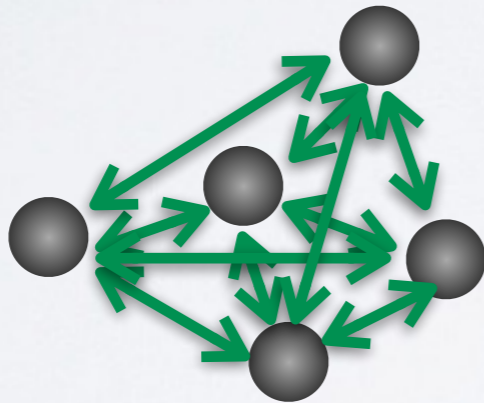
- ▶ Divide **time** into discrete ( $\sim 1$ fs) **time steps** ( $\Delta t$ )  
(for integrating equations of motion, see below)



- ▶ Divide **time** into discrete ( $\sim 1$ fs) **time steps** ( $\Delta t$ )  
(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*

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

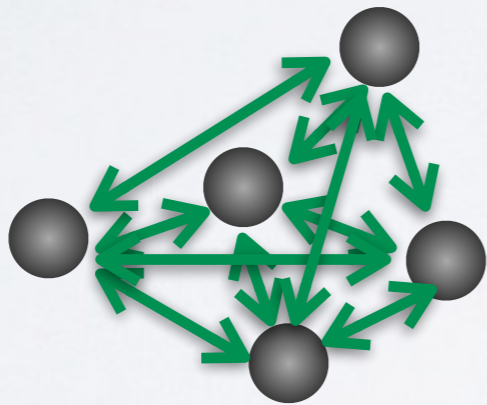
*Empirical force field*

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

- ▶ Divide **time** into discrete ( $\sim 1$ fs) **time steps** ( $\Delta t$ )  
(for integrating equations of motion, see below)



- ▶ At each time step calculate pair-wise atomic **forces** ( $\mathbf{F}(t)$ )  
(by evaluating **force-field** gradient)



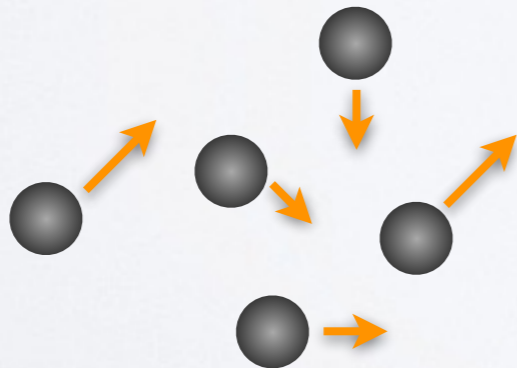
*Nucleic motion described classically*

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

*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)



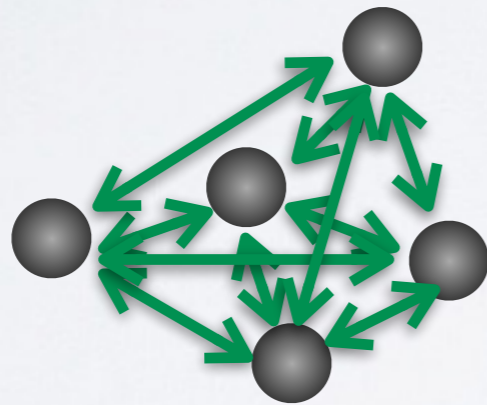
$$\begin{aligned} \mathbf{v}\left(t + \frac{\Delta t}{2}\right) &= \mathbf{v}\left(t - \frac{\Delta t}{2}\right) + \frac{\mathbf{F}(t)}{m} \Delta t \\ \mathbf{r}(t + \Delta t) &= \mathbf{r}(t) + \mathbf{v}\left(t + \frac{\Delta t}{2}\right) \Delta t \end{aligned}$$

# BASIC ANATOMY OF A MD SIMULATION

- ▶ Divide **time** into discrete ( $\sim 1$ fs) **time steps** ( $\Delta t$ )  
(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*

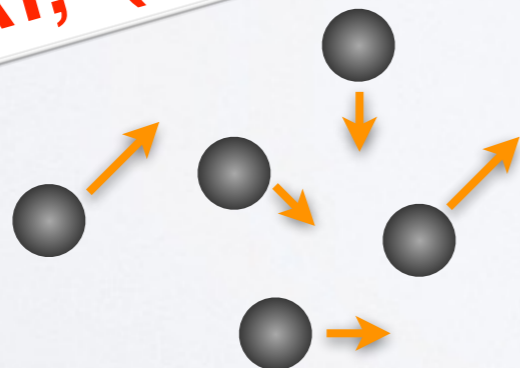
$$m_i \frac{d^2 \vec{R}_i}{dt^2} = -\vec{\nabla}_i E(\vec{R})$$

*Empirical force field*

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

- ▶ Use the forces to calculate **velocities** and move atoms to new **positions**  
(the integration is done numerically via the “leapfrog” scheme)

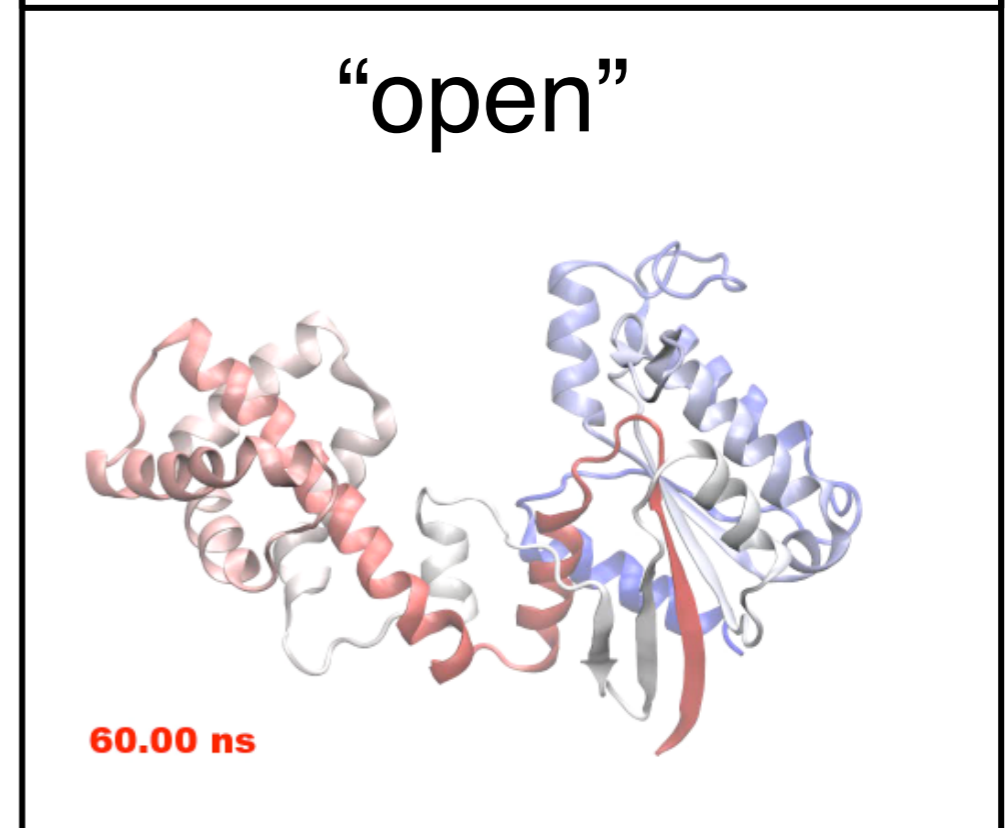
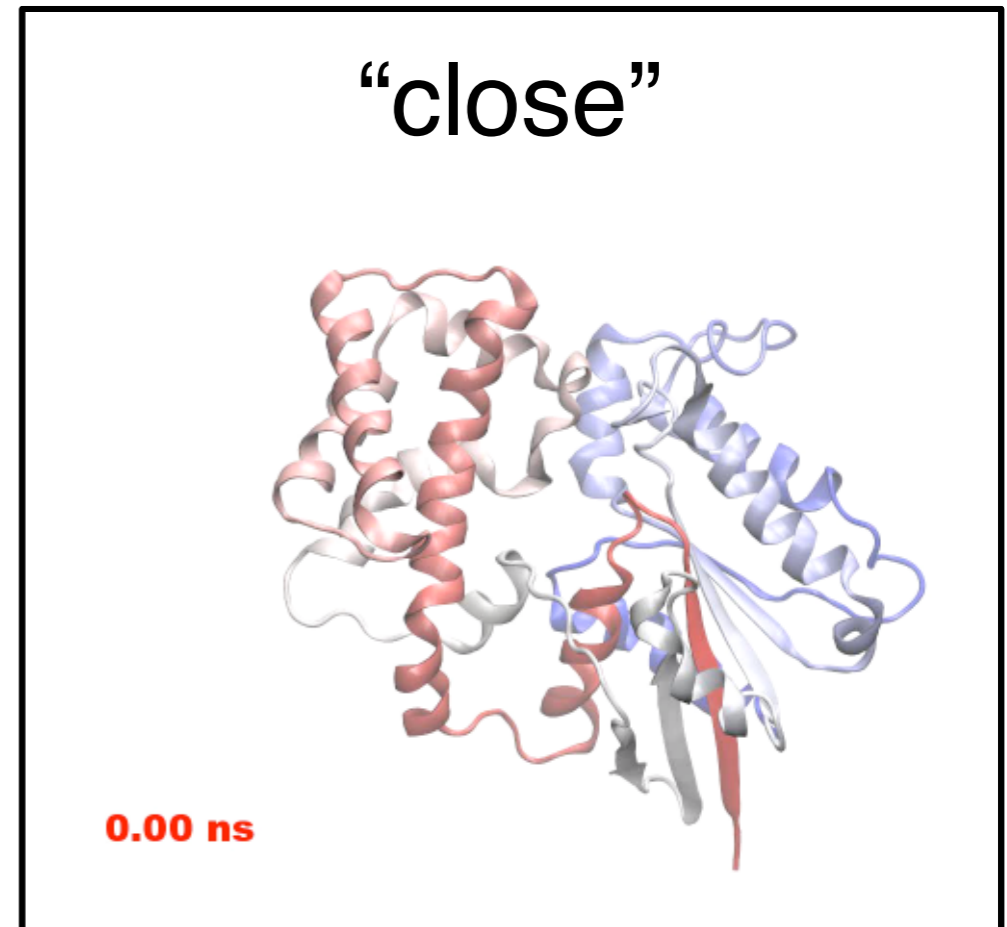
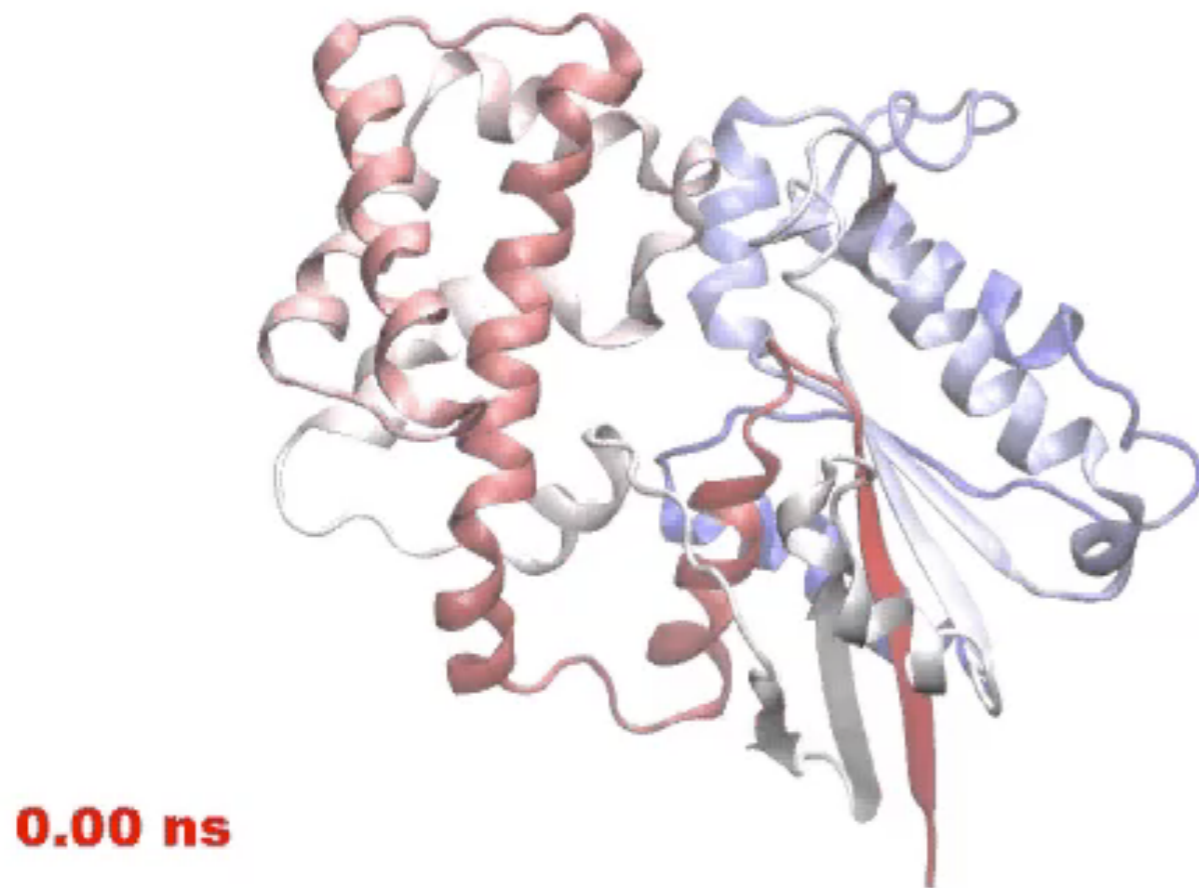
**REPEAT, (iterate many, many times... 1ms = 10<sup>12</sup> time steps)**



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

# MD Prediction of Functional Motions

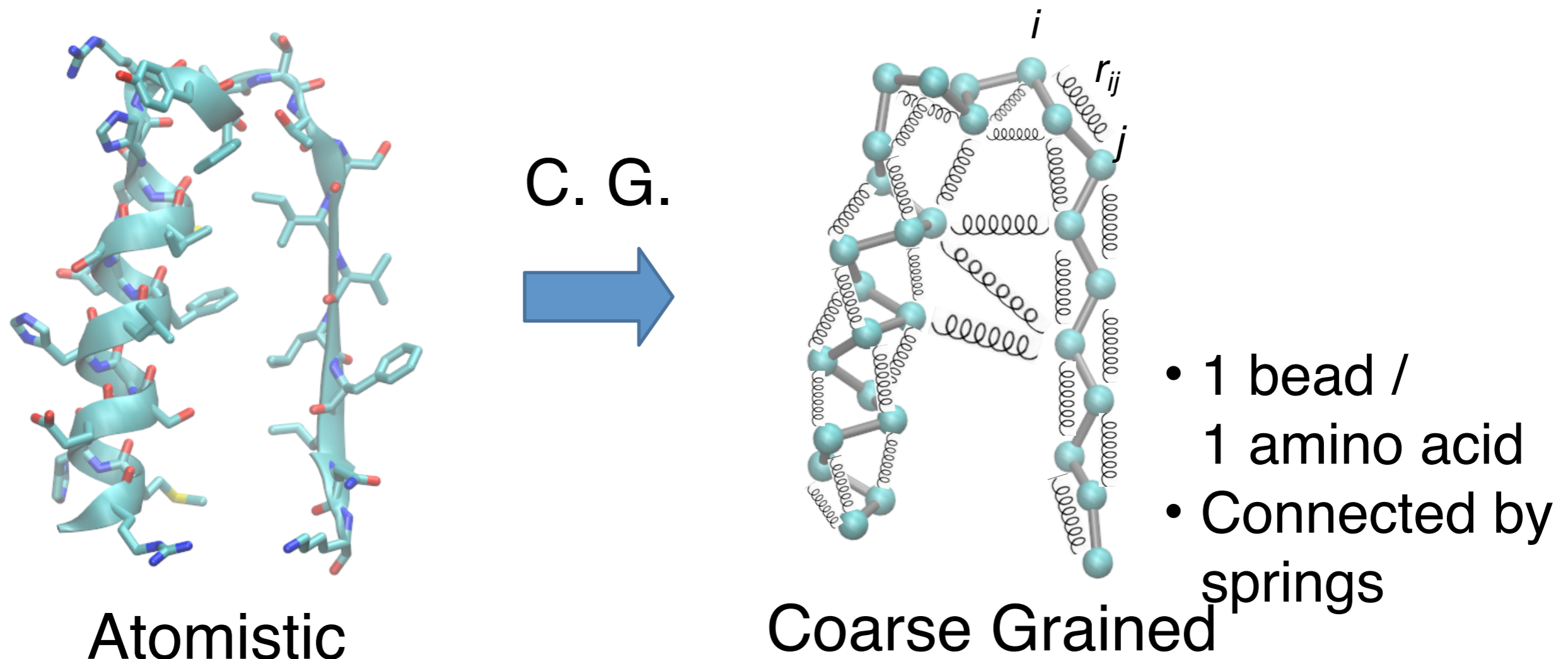
Accelerated MD simulation of  
nucleotide-free transducin alpha subunit



Yao and Grant, Biophys J. (2013)

# 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.





## ACHIEVEMENTS

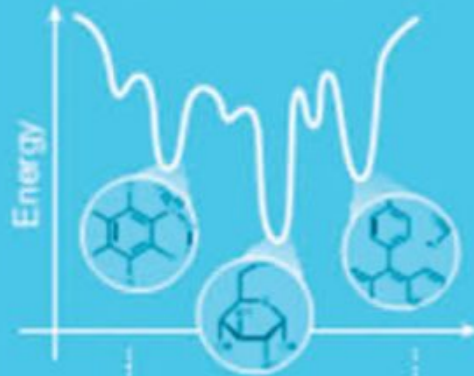
Computational power



Data coverage and community resources



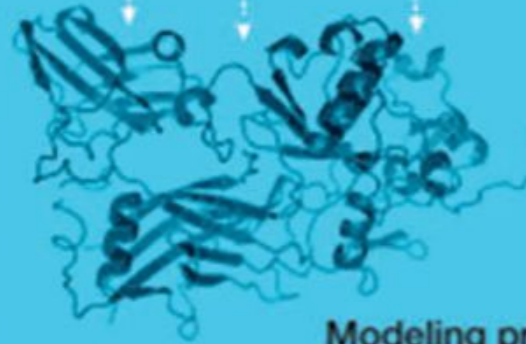
Chemical systems biology and small-molecule docking simulations



Objective method assessment



Correlated mutations



Modeling protein structure

## CHALLENGES

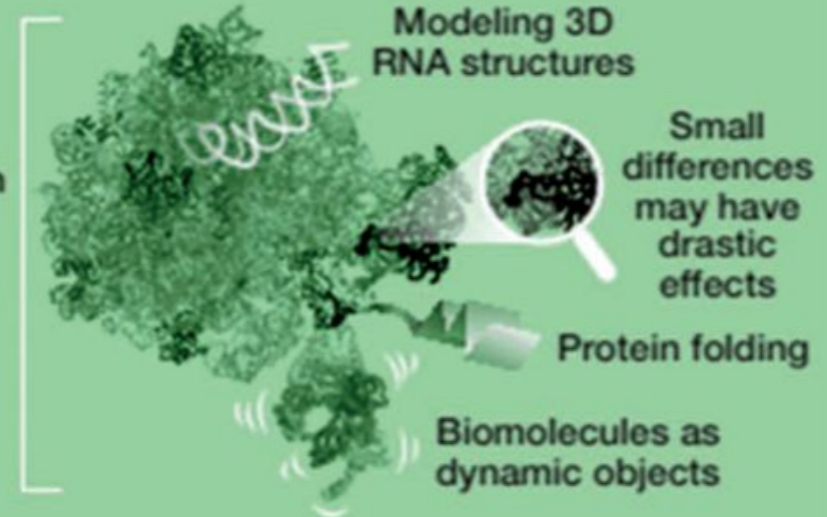
Accessibility and integration of data and methods



Protein engineering and synthetic biology



Modeling multi-domain proteins and large assemblies



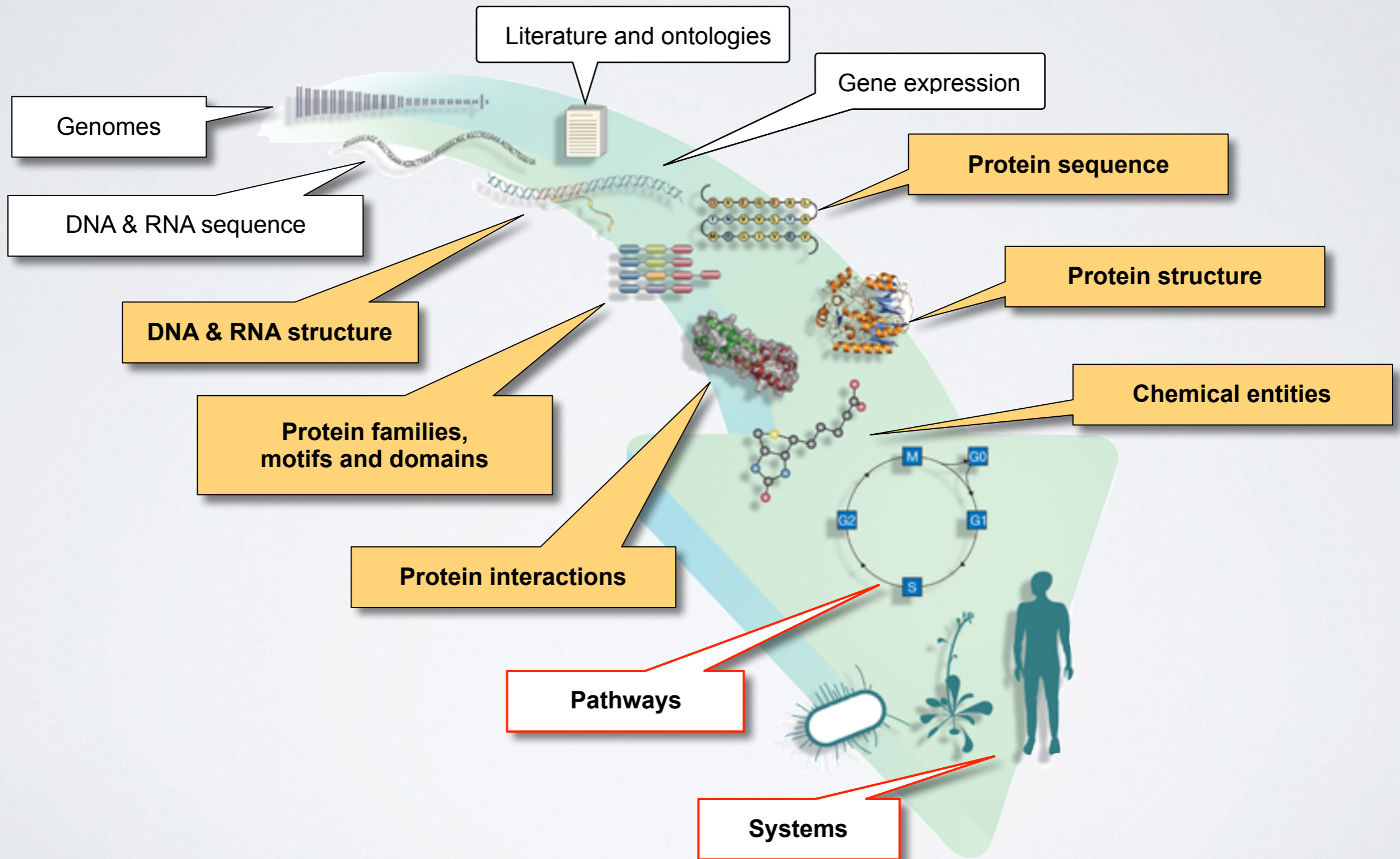
Origins and evolution of protein structure



Integration with systems biology



# 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
- Explored how to use R to perform structural bioinformatics analysis!
- Introduced both physics and knowledge based modeling approaches for describing the structure, energetics and dynamics of proteins computationally
- Introduced both structure and ligand based bioinformatics approaches for drug discovery and design

[ [Muddy Point Assessment](#) ]

# CAUTIONARY NOTES

- **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 parameters is an ongoing 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.