

BGGN 213

Structural Bioinformatics II

Lecture 12

Barry Grant
UC San Diego

<http://thegrantlab.org/bggn213>

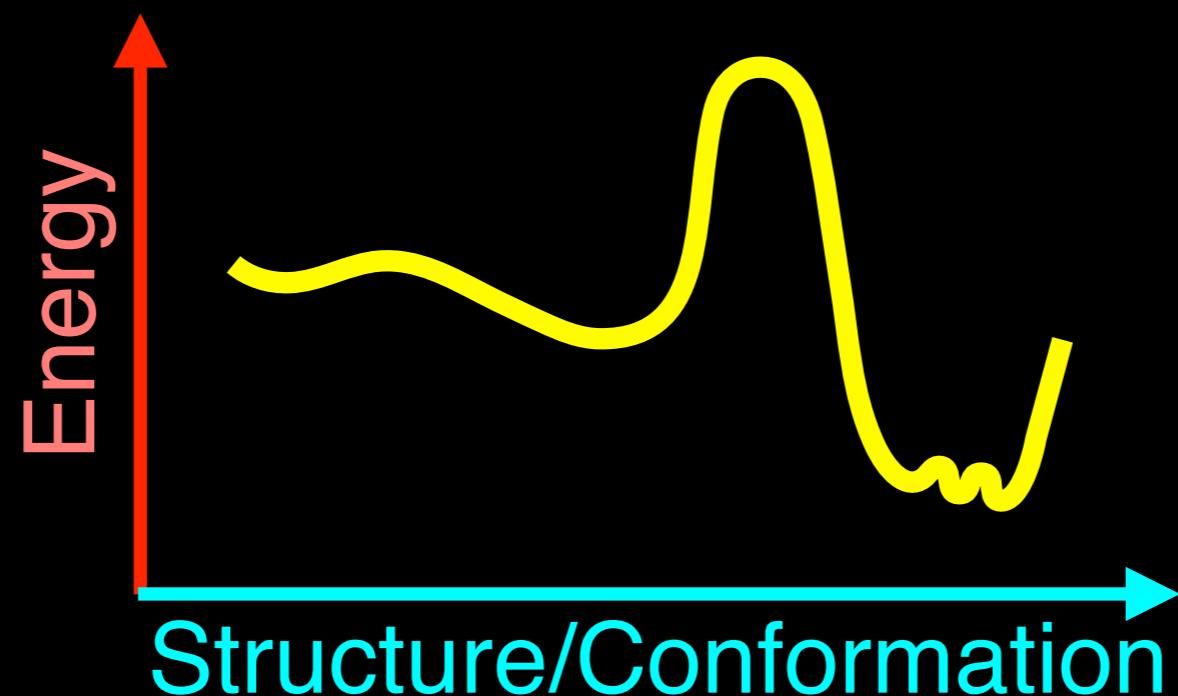
Download MGL Tools: See class website!

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 systems
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

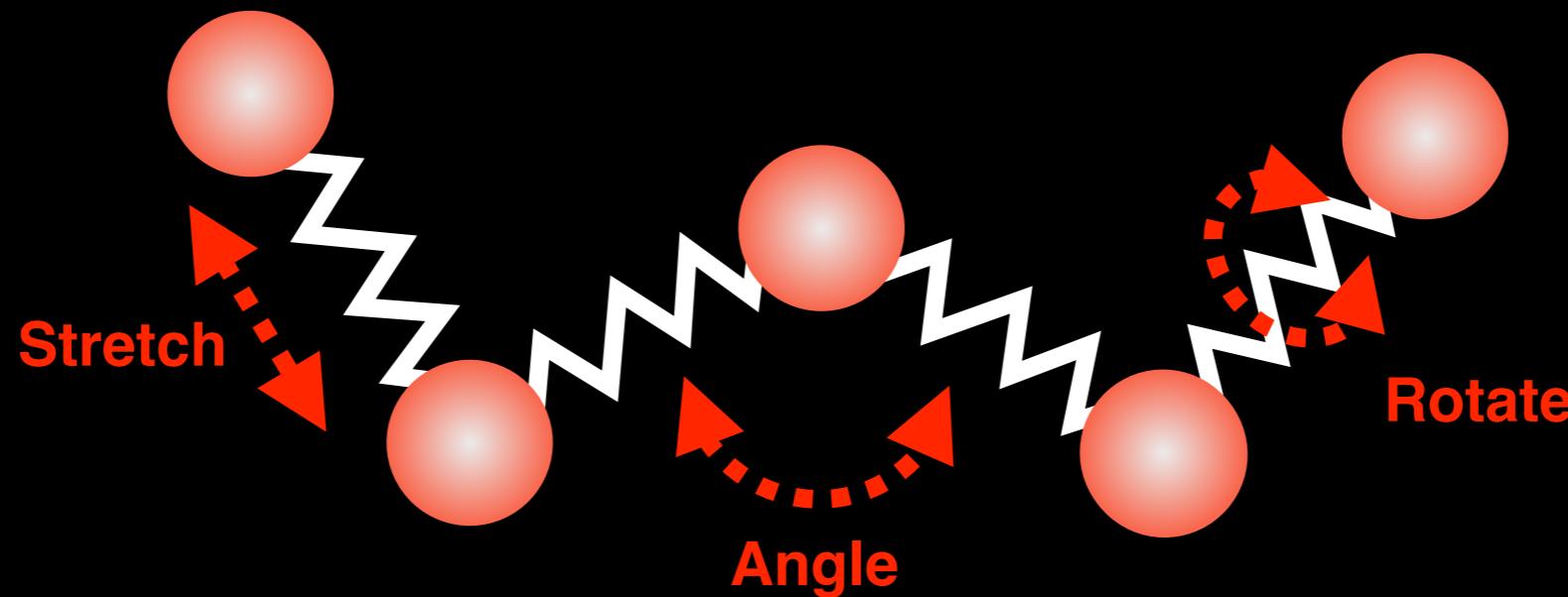
$$V(R) = E_{\text{bonded}} + E_{\text{non.bonded}}$$

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

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

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

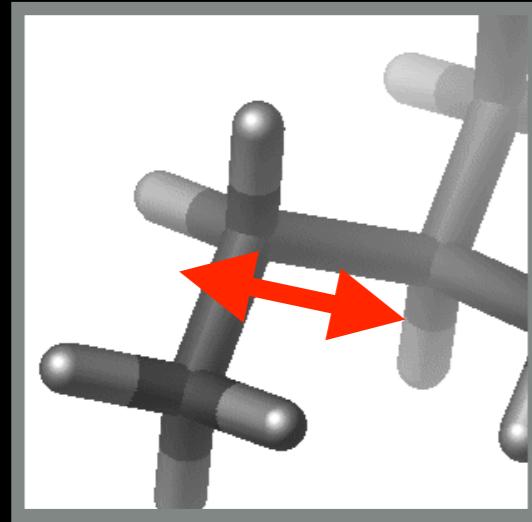
$$E_{\text{bond.stretch}} + E_{\text{bond.angle}} + E_{\text{bond.rotate}}$$



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

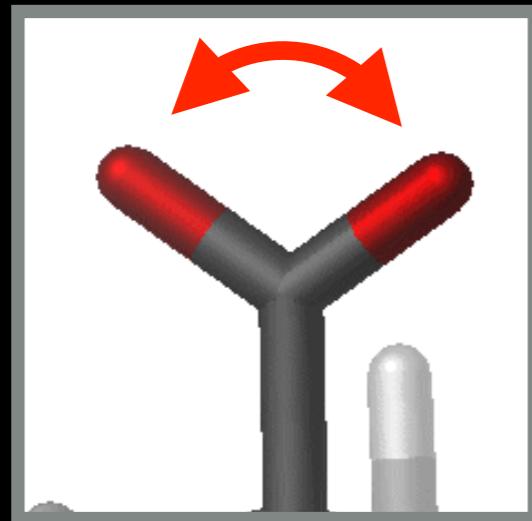
E_{bonded} is itself a sum of three terms:

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



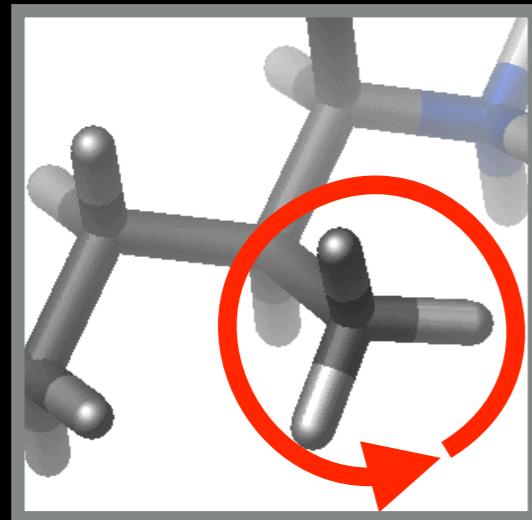
Bond Stretch

$E_{bond.stretch}$



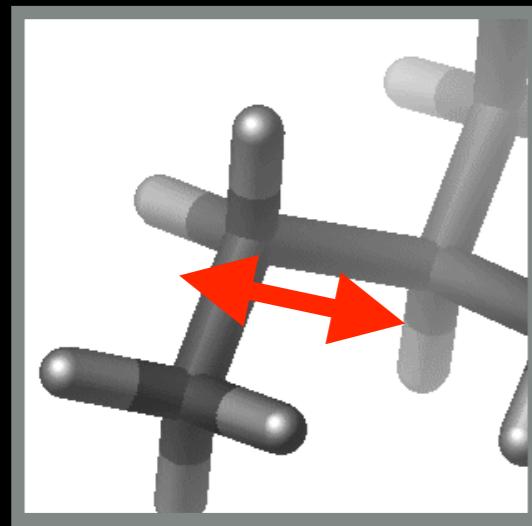
Bond Angle

$E_{bond.angle}$



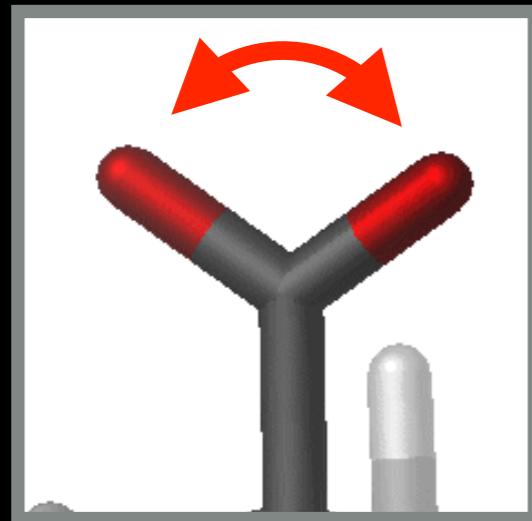
Bond Rotate

$E_{bond.rotate}$



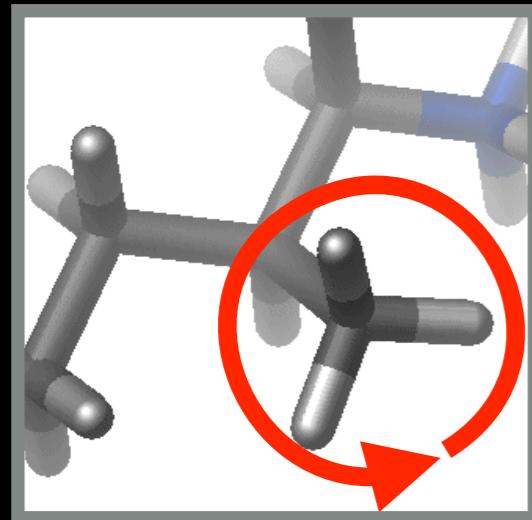
Bond Stretch

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



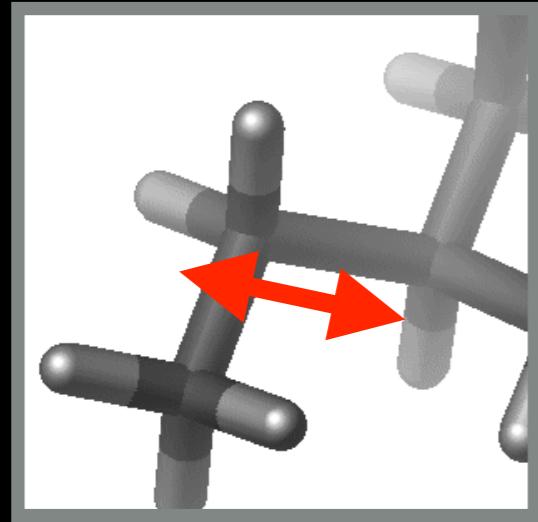
Bond Angle

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



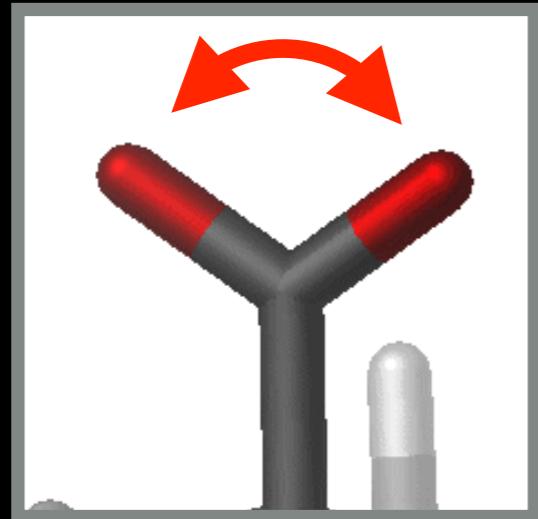
Bond Rotate

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



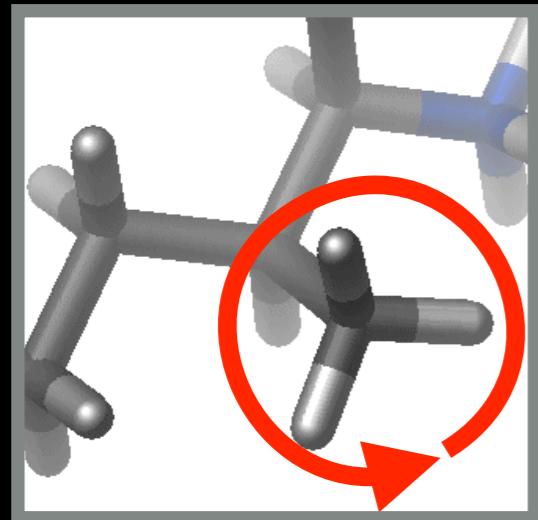
Bond Stretch

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



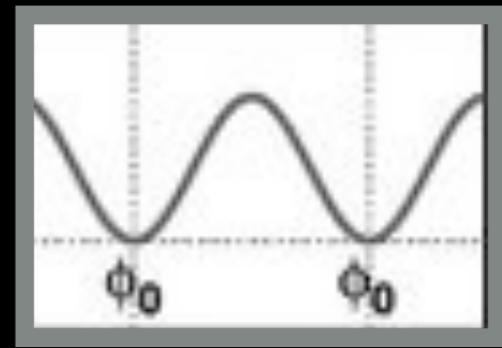
Bond Angle

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



Bond Rotate

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



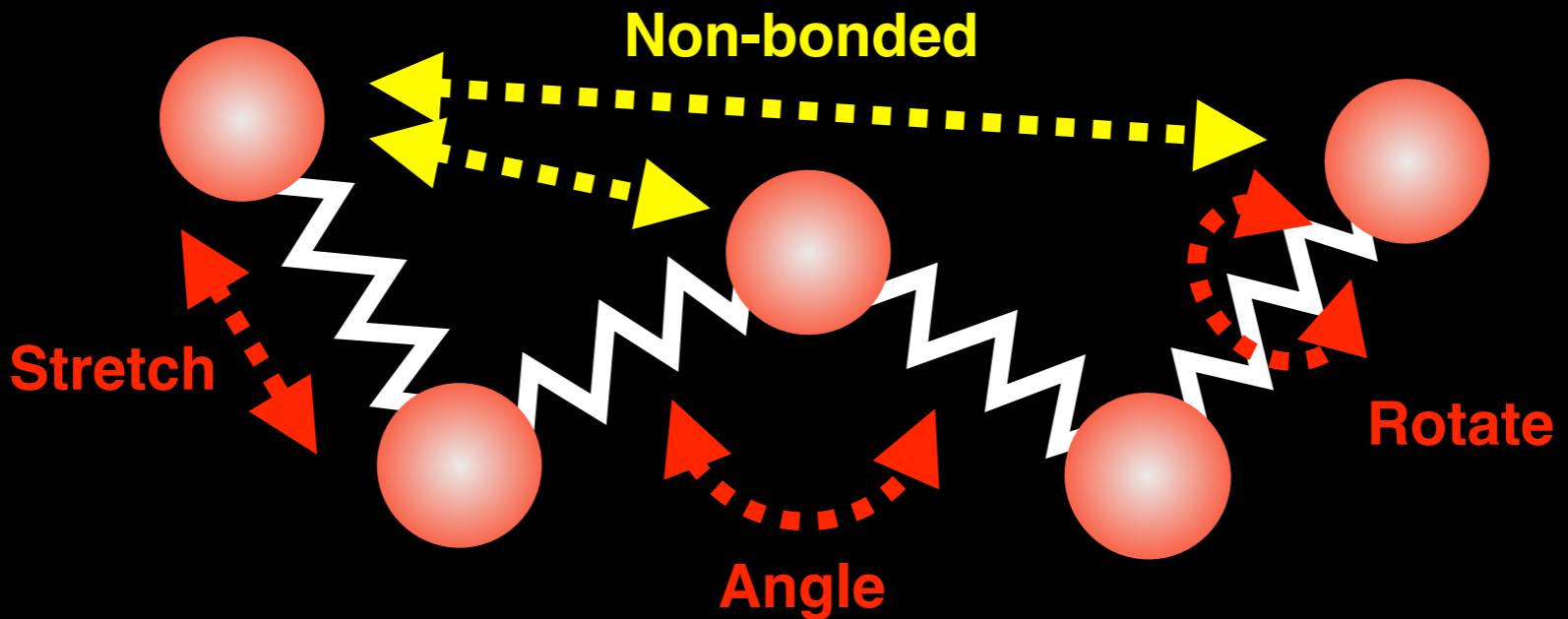
$$V(R) = E_{bonded} + \boxed{E_{non.bonded}}$$

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

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

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

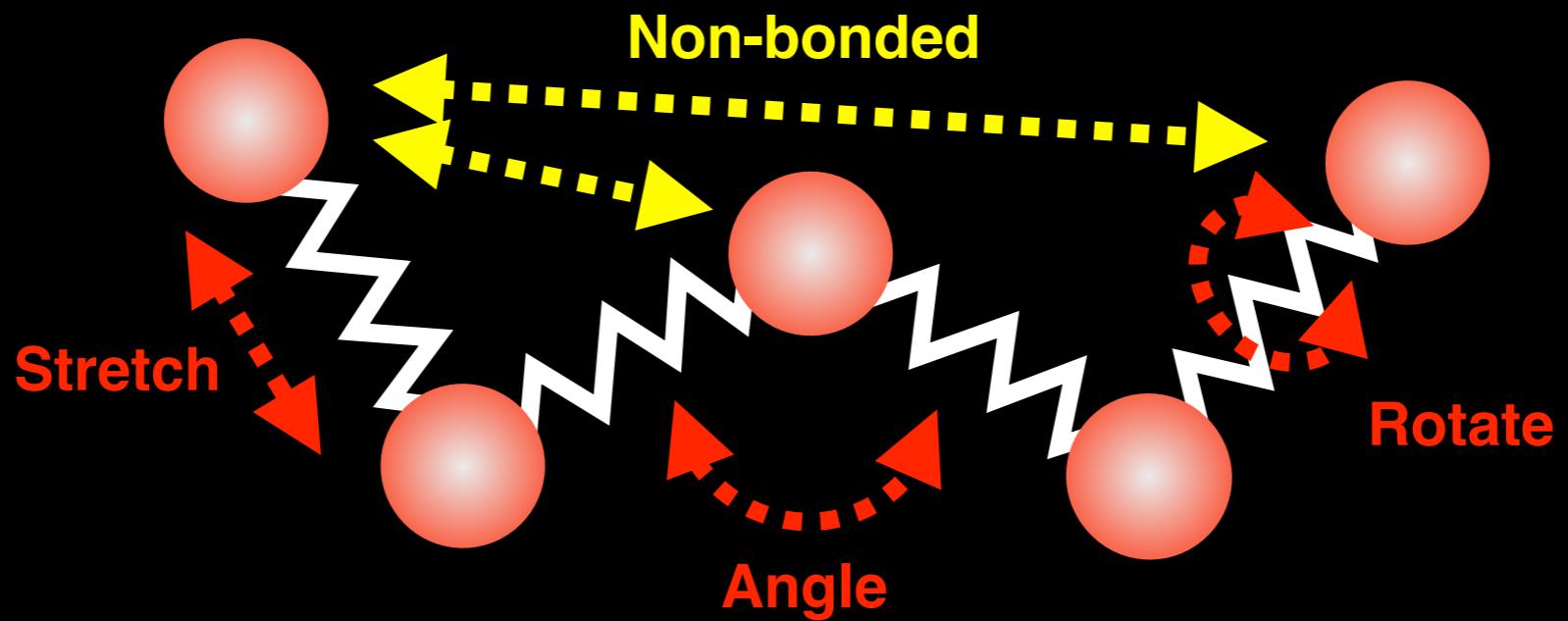
$$E_{van.der.Waals} + E_{electrostatic}$$



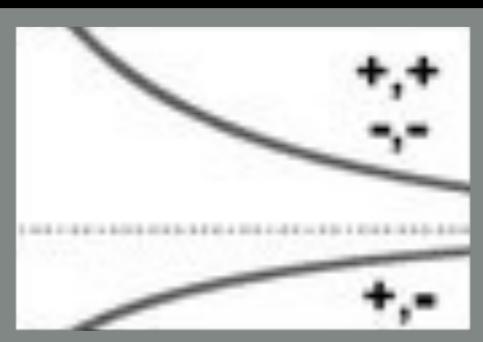
$$V(R) = E_{bonded} + \boxed{E_{non.bonded}}$$

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

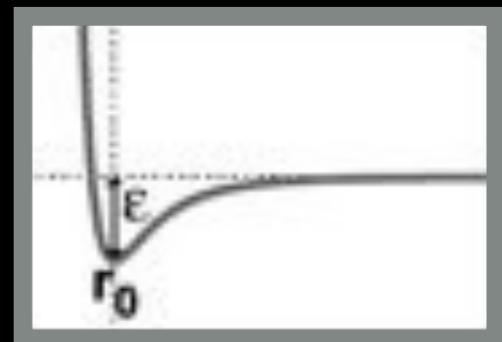
$$E_{van.der.Waals} + E_{electrostatic}$$



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



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



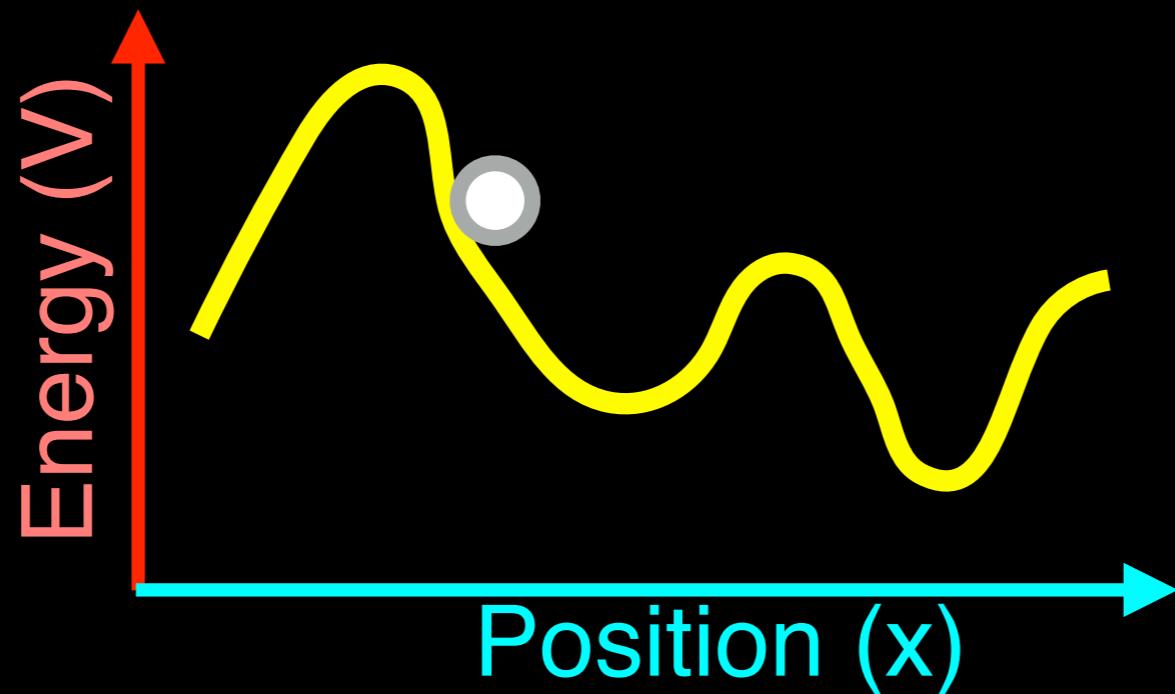
Total potential energy

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

$$V(R) = E_{\text{bonded}} + E_{\text{non.bonded}}$$
$$\left. \begin{aligned} &E_{\text{bond.stretch}} \\ &+ E_{\text{bond.angle}} \\ &+ E_{\text{bond.rotate}} \\ &+ E_{\text{van.der.Waals}} \\ &+ E_{\text{electrostatic}} \end{aligned} \right\} E_{\text{bonded}}$$
$$\left. \begin{aligned} & \\ & \\ & \\ & \\ & \end{aligned} \right\} E_{\text{non.bonded}}$$

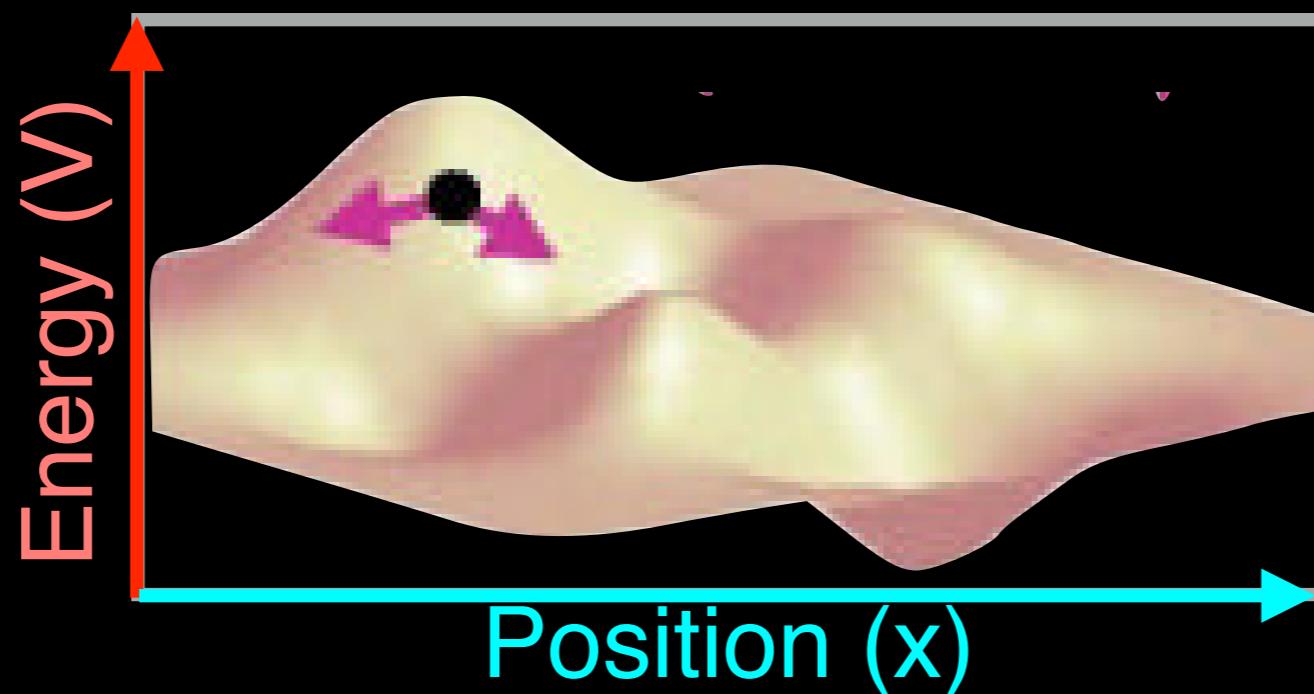
Potential energy surface

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



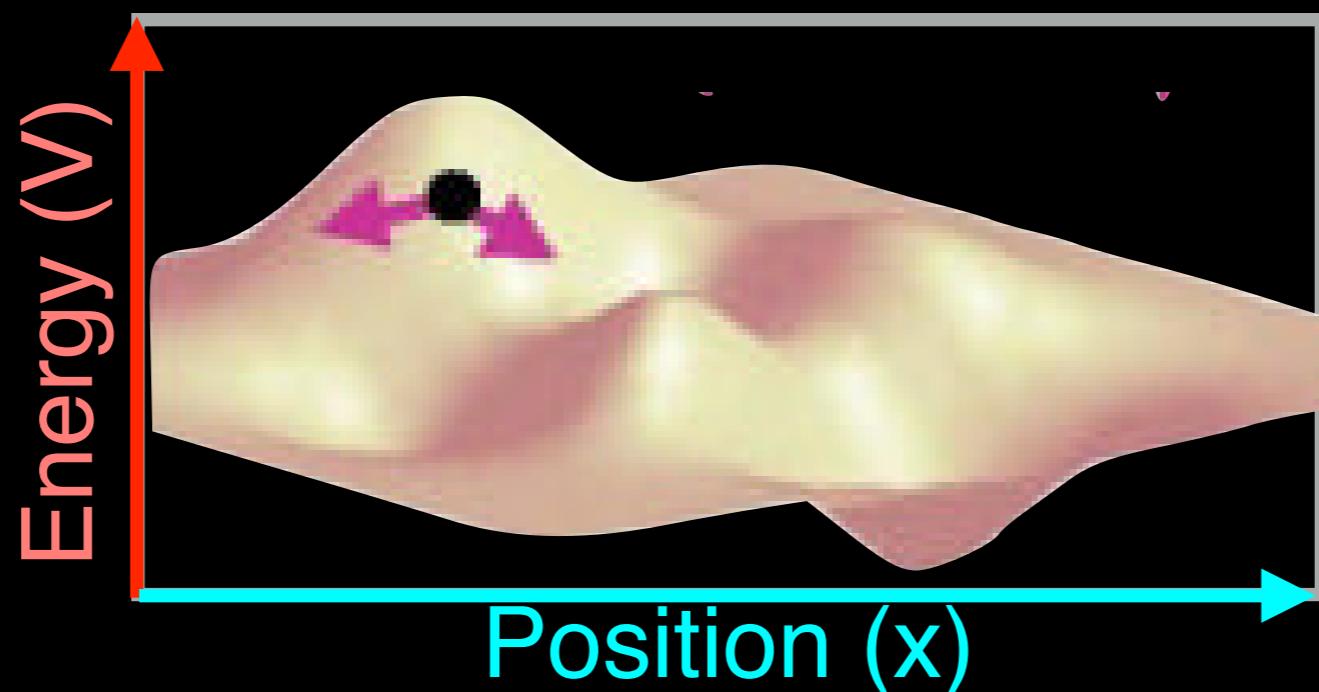
Potential energy surface

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



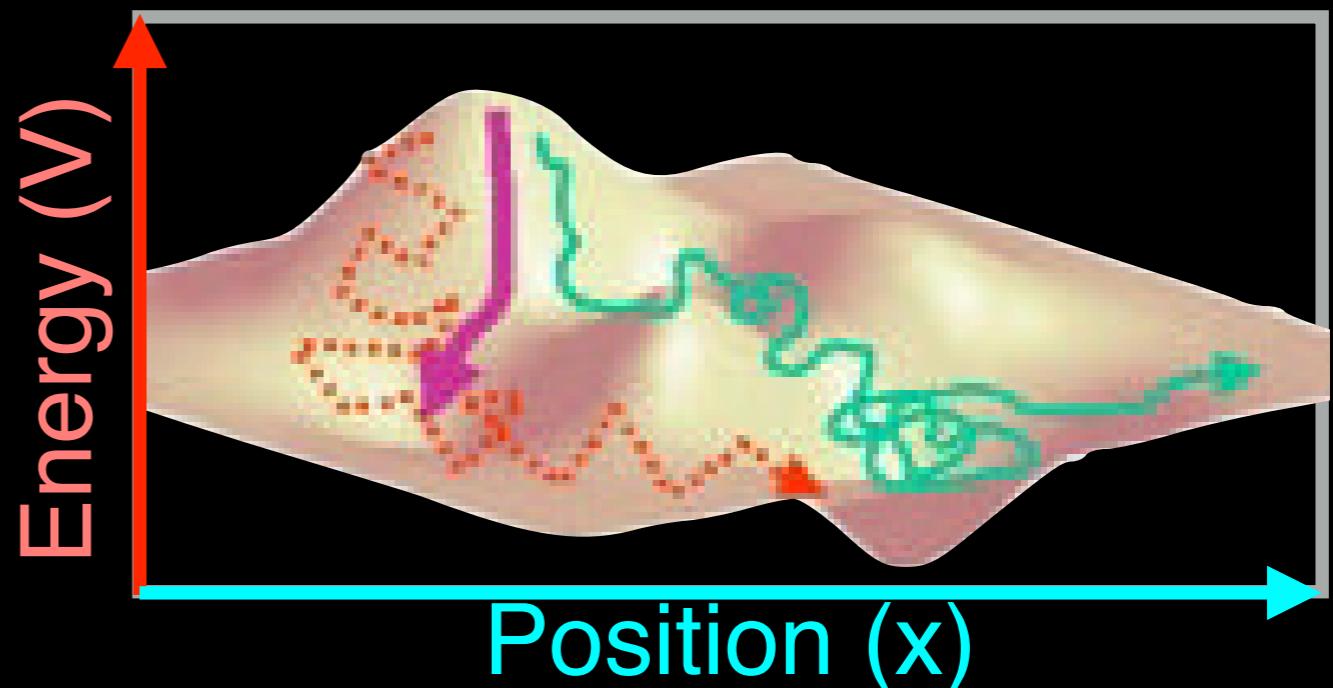
- The **forces** are the gradients of the energy

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

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

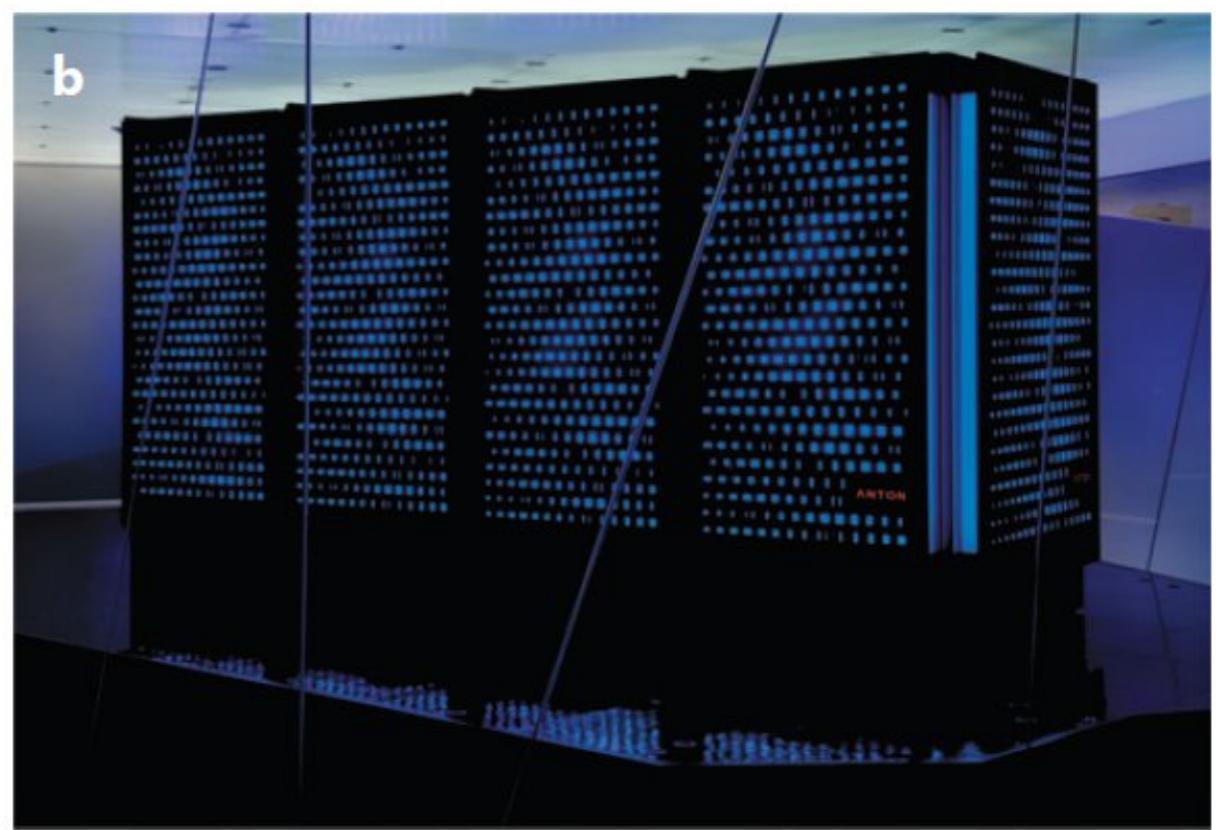
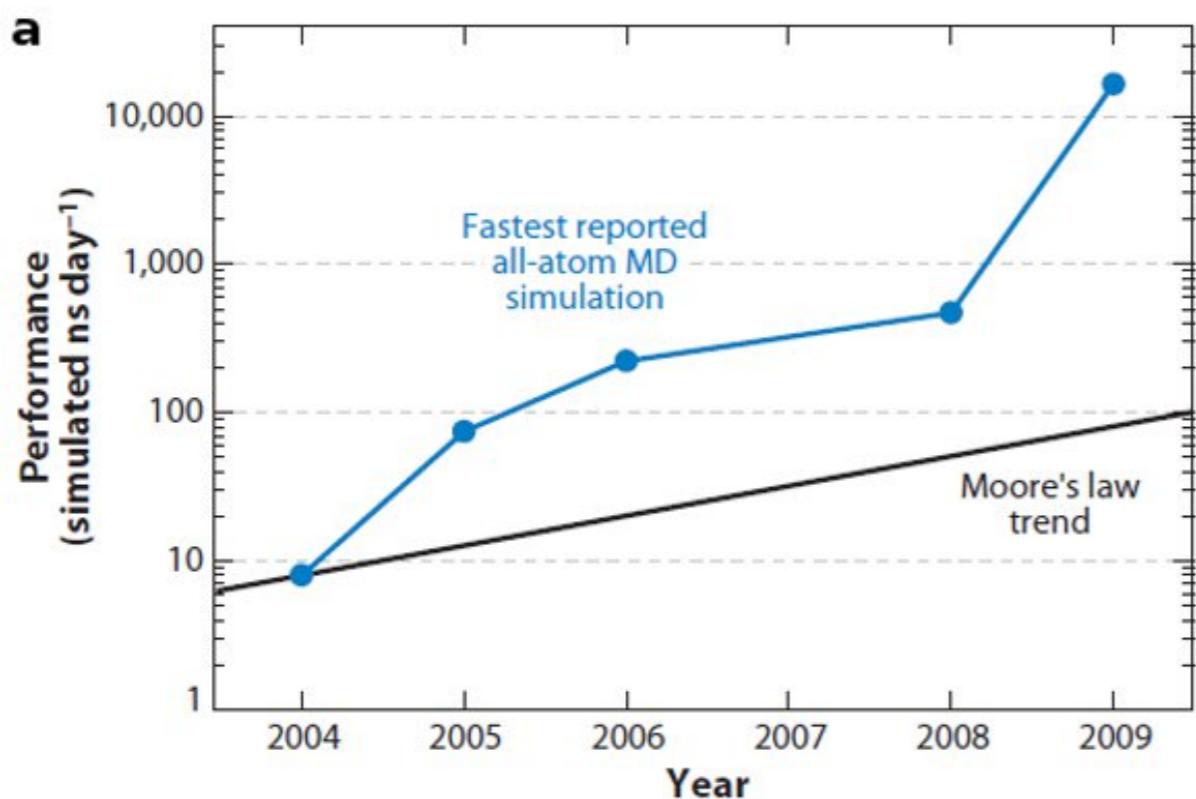
HOW COMPUTERS HAVE CHANGED

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

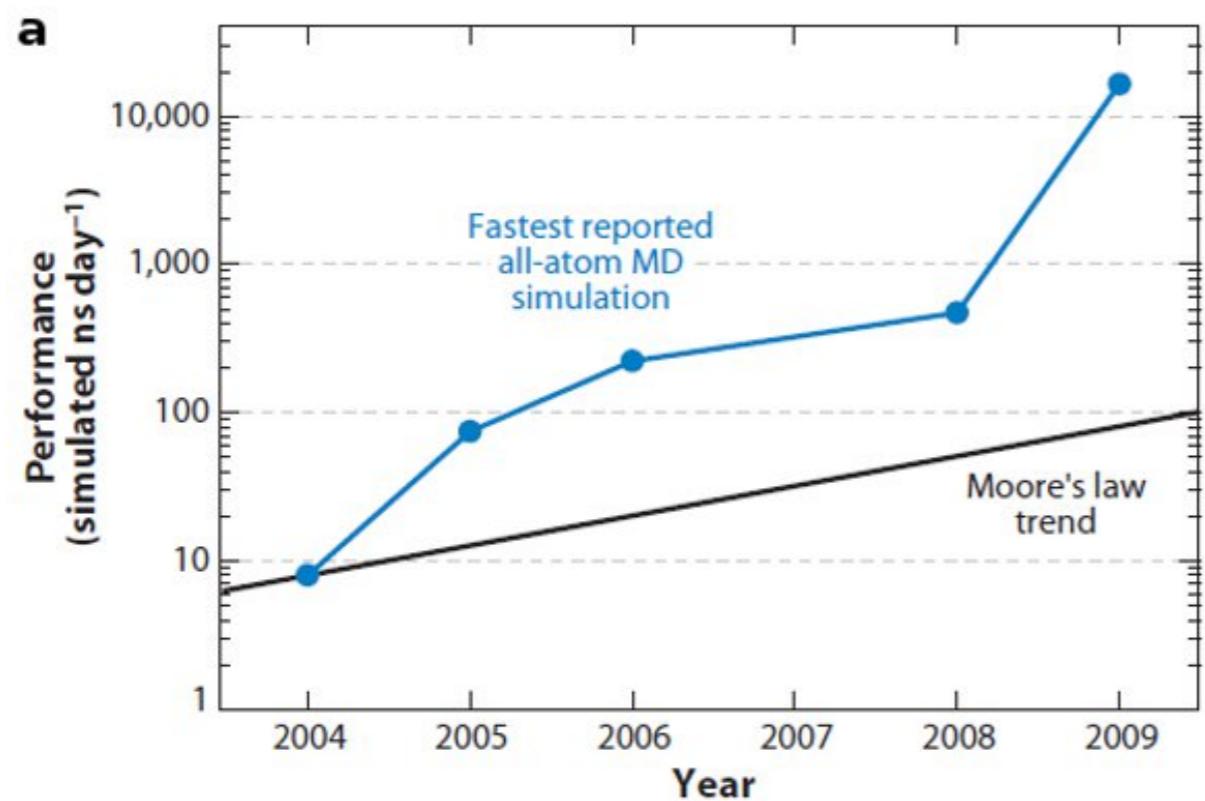
If cars were like computers then a new Volvo would cost \$3, would have a top speed of 1,000,000 Km/hr, would carry 50,000 adults and would park in a shadow.



SIDE-NOTE: GPUS AND ANTON SUPERCOMPUTER



SIDE-NOTE: GPUS AND ANTON SUPERCOMPUTER

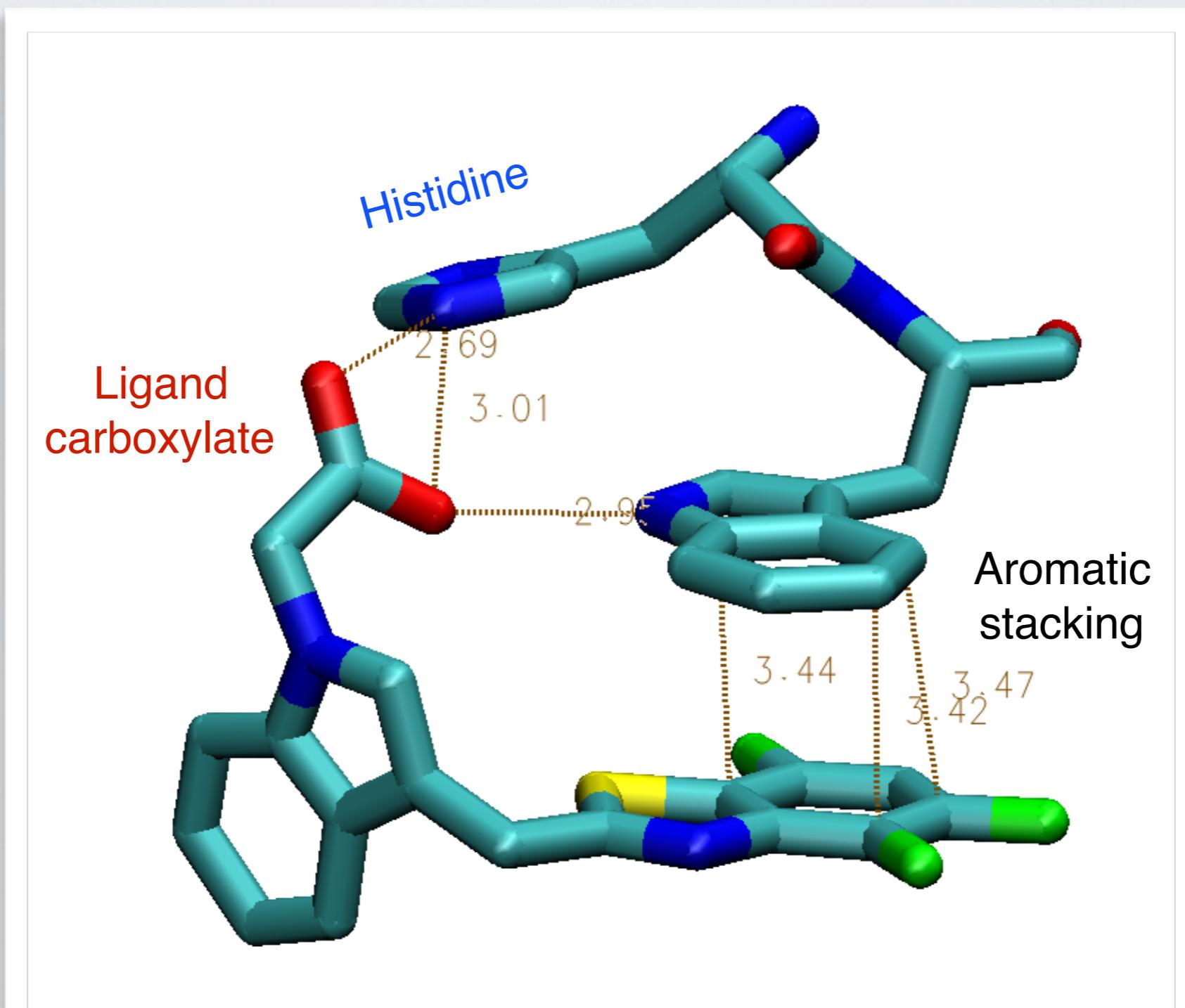


POTENTIAL FUNCTIONS DESCRIBE A SYSTEMS **ENERGY** AS A FUNCTION OF ITS **STRUCTURE**

Two main approaches:

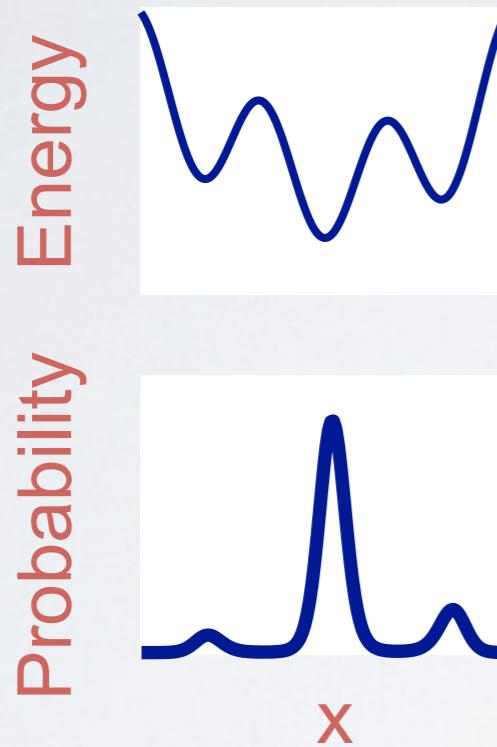
- (1). Physics-Based
- (2). Knowledge-Based

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 $p(r_{O-N})$
3. Compute $E(r_{O-N})$ from $p(r_{O-N})$

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

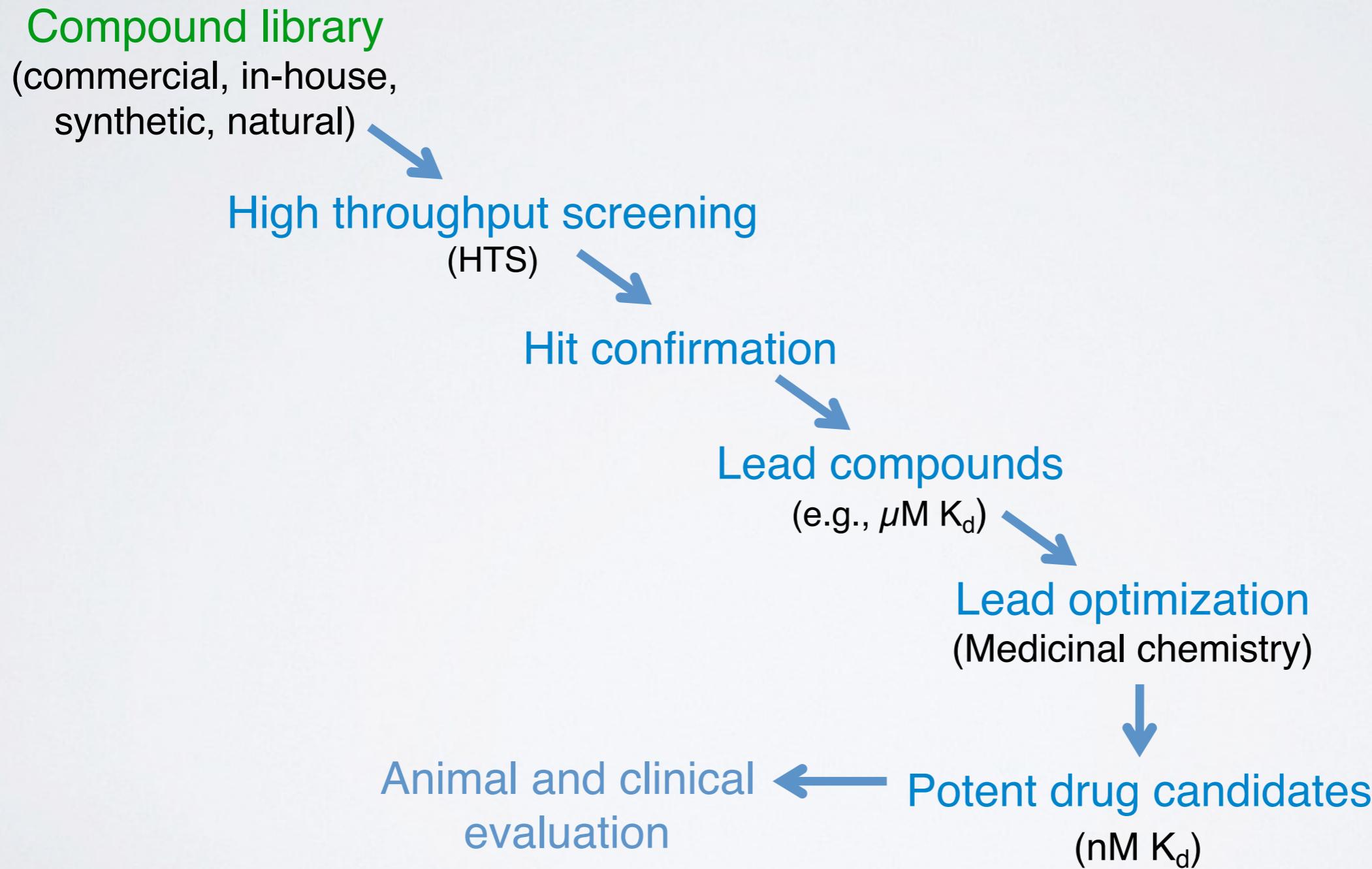
(not always clear how to make improvements)

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



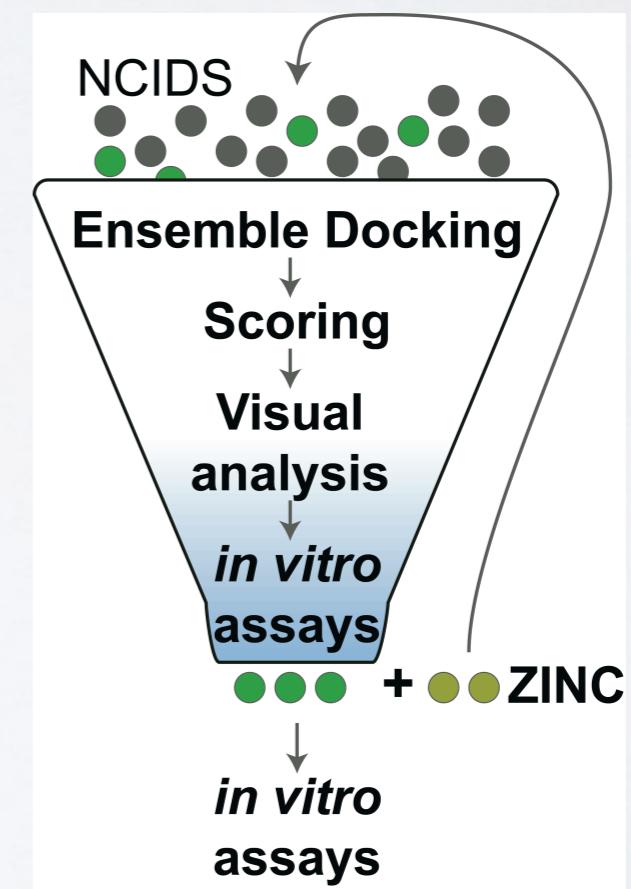
COMPUTER-AIDED LIGAND DESIGN

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:

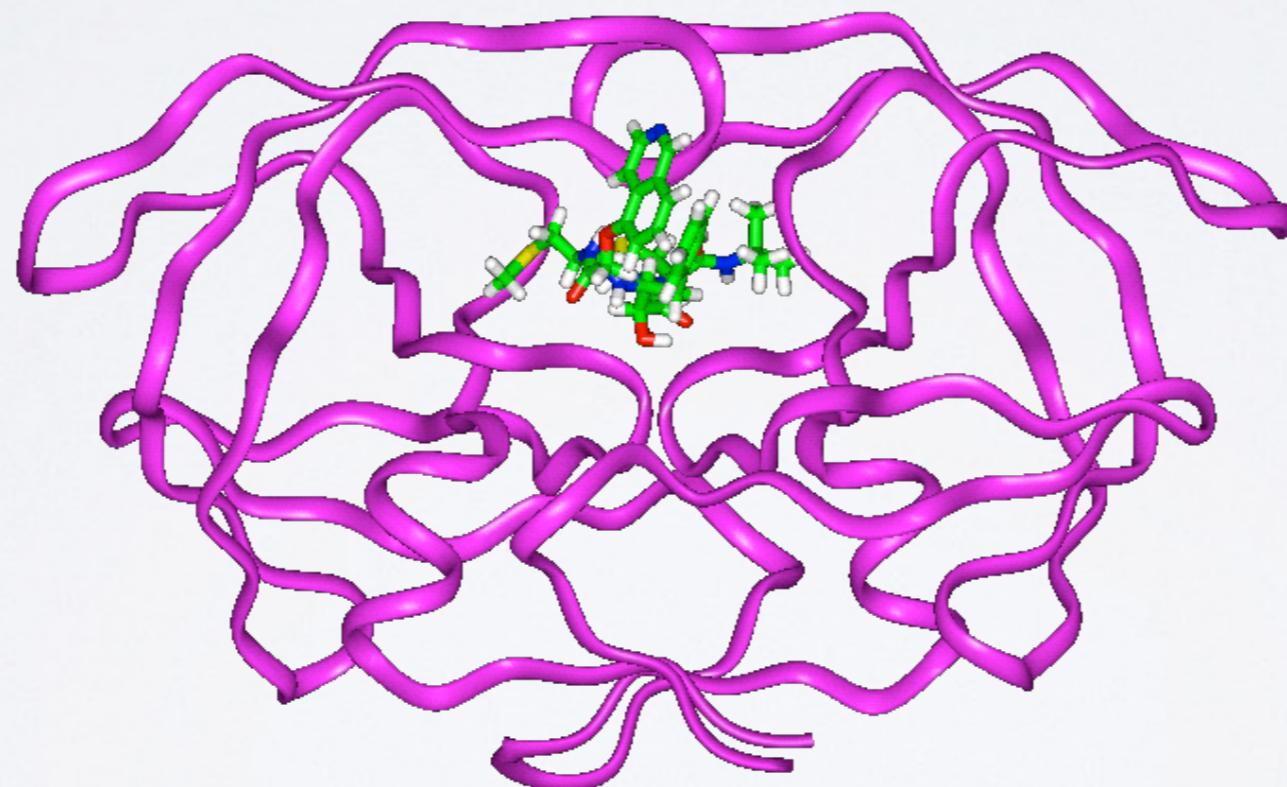
(1). Receptor/Target-Based

(2). Ligand/Drug-Based

SCENARIO I:

RECEPTOR-BASED DRUG DISCOVERY

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



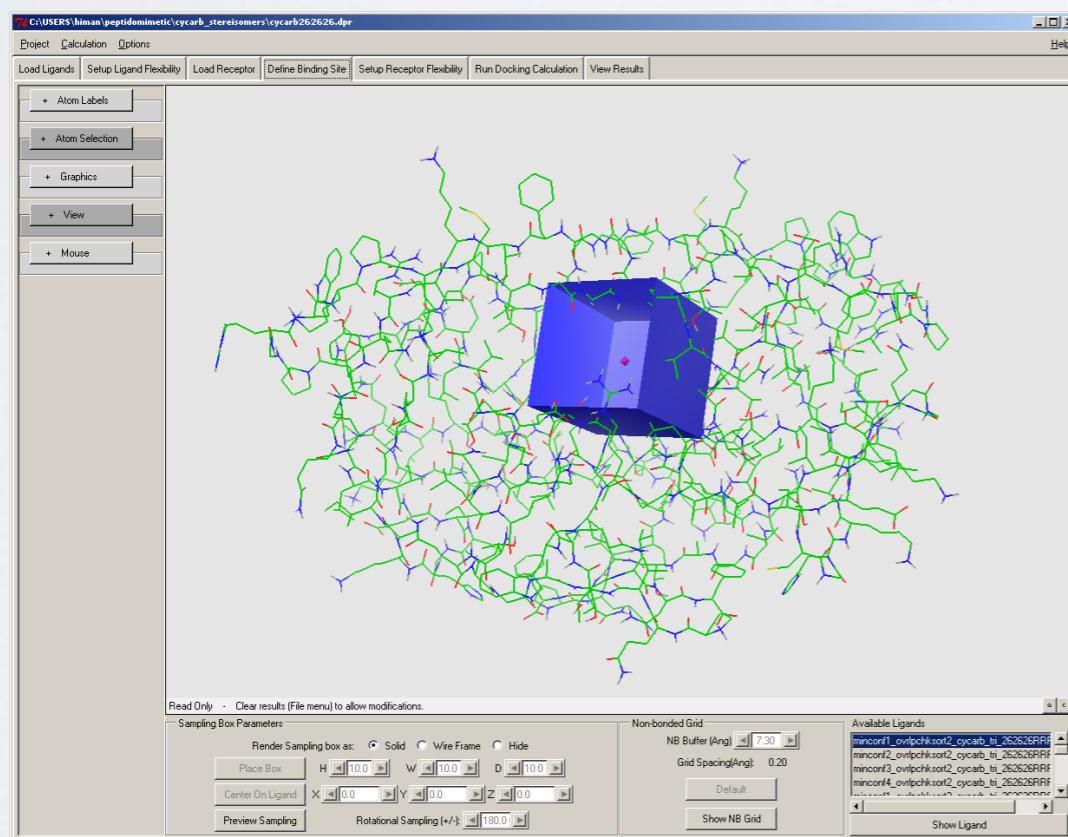
HIV Protease/KNI-272 complex

PROTEIN-LIGAND DOCKING

Structure-Based Ligand Design

Docking software

Search for structure of lowest energy



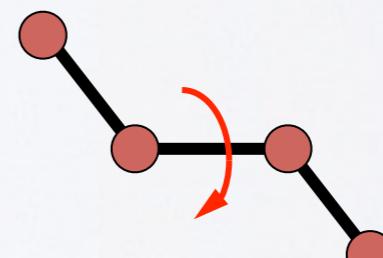
Potential function
Energy as function of structure



VDW

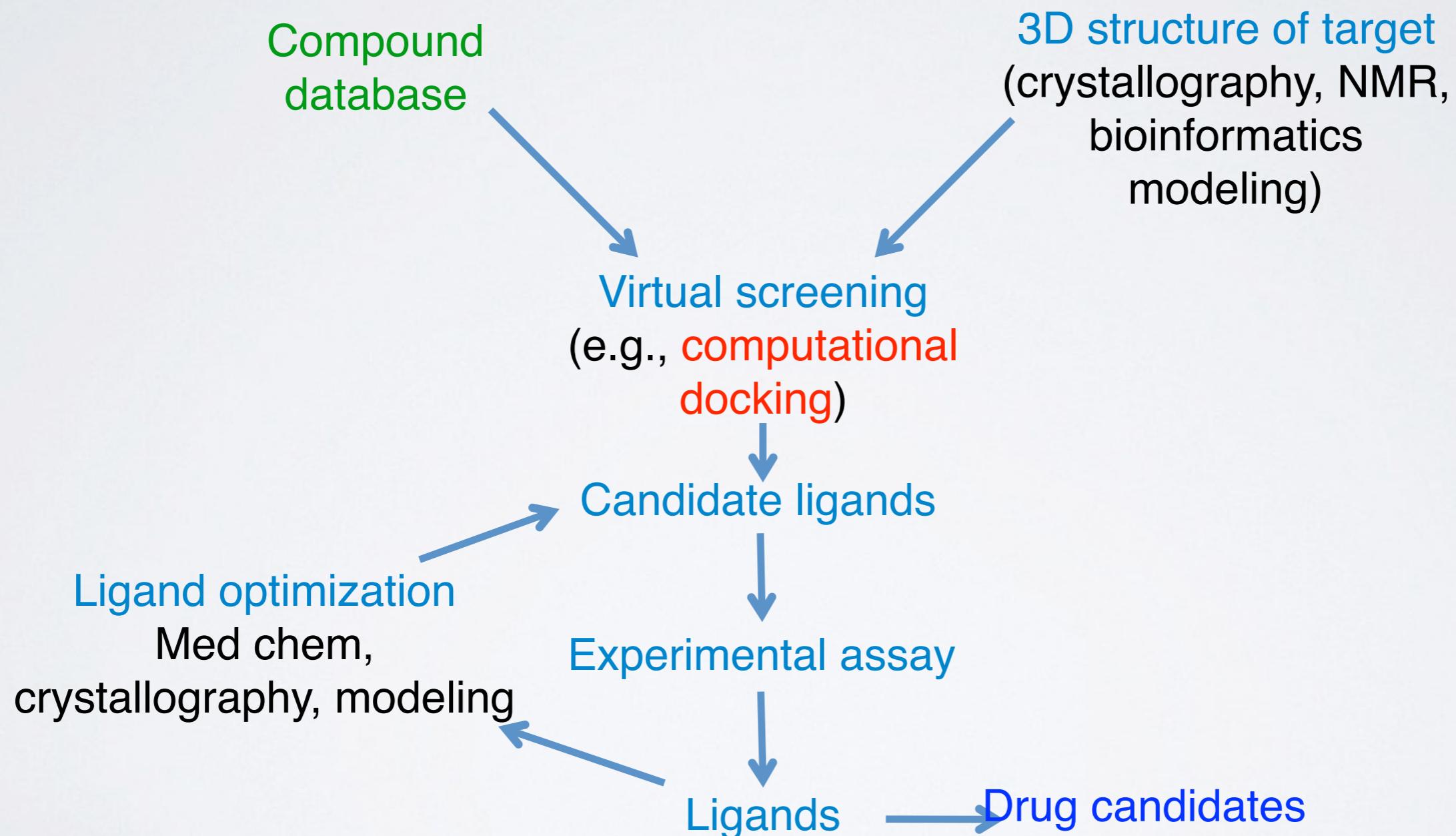


Screened Coulombic



Dihedral

STRUCTURE-BASED VIRTUAL SCREENING



COMPOUND LIBRARIES

MAYBRIDGE
Part of Thermo Fisher Scientific

Search center
 Building Blocks
 Screening
 Select the catalog you want to query
 Go
 Building Blocks
 Screening
 Structure search

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

Maximise quality hits from your screens

- The HitFinder™ Collection comprises 14,400 premier compounds representing the drug-like diversity of the Maybridge Screening Collection, offering easy and rapid lead identification.
- Selections are made using a clustering algorithm employing standard Daylight Fingerprints with the Tanimoto similarity index clustering at 0.71 similarity*.

Reduced time to optimize any hit

- All screening compounds fit Lipinski guidelines for "Drug-likeness**", and all have purity greater than 90%.
- Compounds have been selected to be non-reactive, ensuring fewer false positives and higher quality results.
- When you are ready to optimize your drug lead, our range of over 6000 advanced novel Maybridge Building Blocks gives high chemical diversity for accelerating your drug design process.

Ready to Screen



All HitFinder™ plates are securely sealed and carry both a clear plate number and bar-code for convenient use and storage.

To download the HitFinder™ database files [click here](#).
 To download the PDF file [click here](#).

NIH MOLECULAR LIBRARIES
 SMALL MOLECULE REPOSITORY

A Roadmap Initiative

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.

Registered Users Login

Copyright © 2007 Galapagos NV

BioFocus
 A Galapagos Company

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

PITT HOME | FIND PEOPLE | CONTACT US

University of Pittsburgh
Pittsburgh Molecular Libraries Screening Center

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
CONTACTS

 Go

PMLSC
BIG DISCOVERIES FROM SMALL MOLECULES

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](#)
 © Office of the Senior Vice Chancellor for Health Sciences, University of Pittsburgh. All rights reserved.
 Last Update 3/14/2007

Commercial
 (in-house pharma)

Government (NIH)

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/bggn213_S19/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**

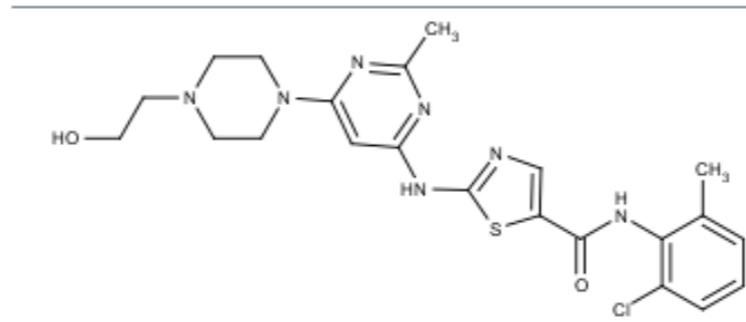
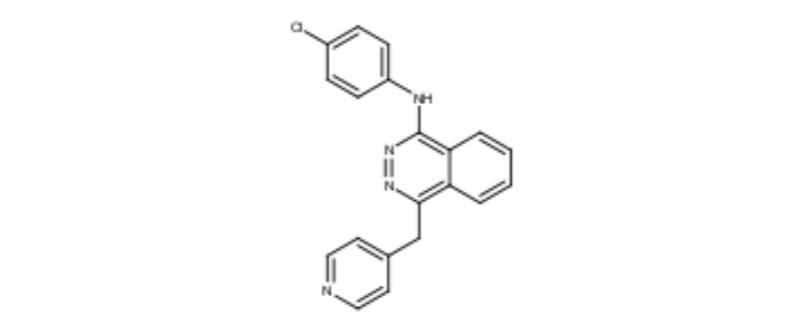
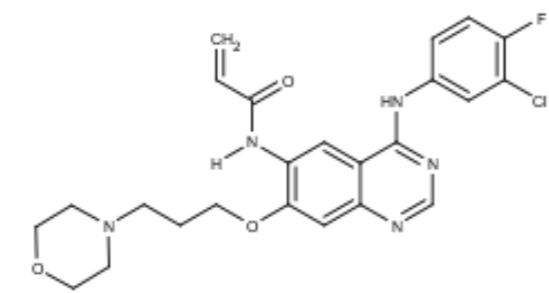
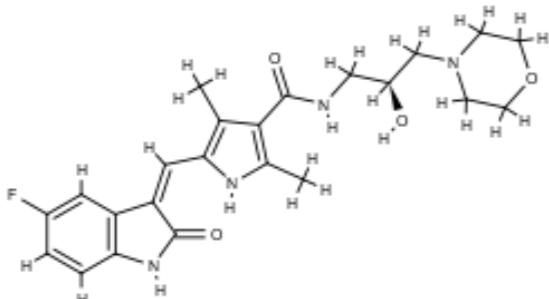
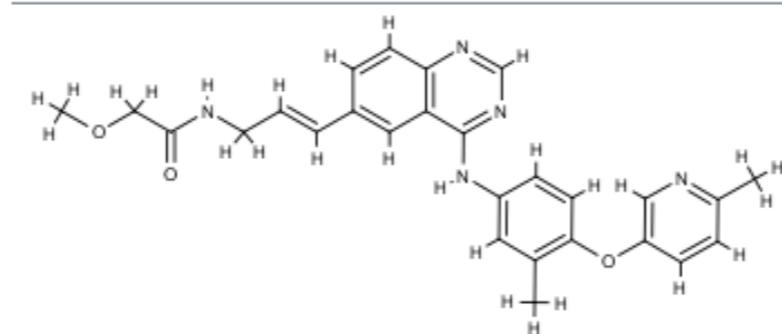
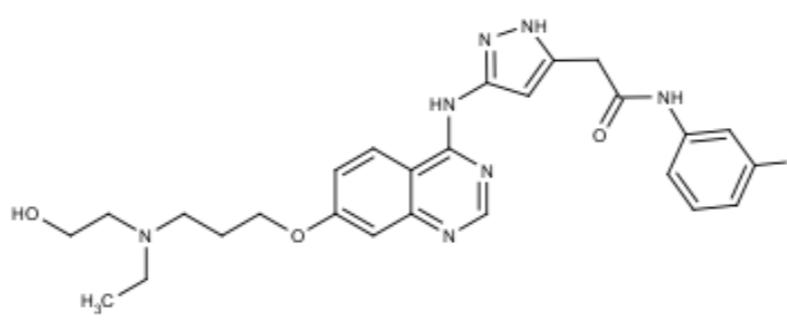
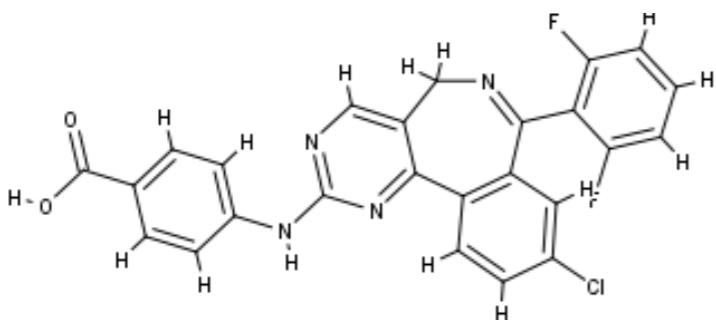
Two main approaches:

- (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

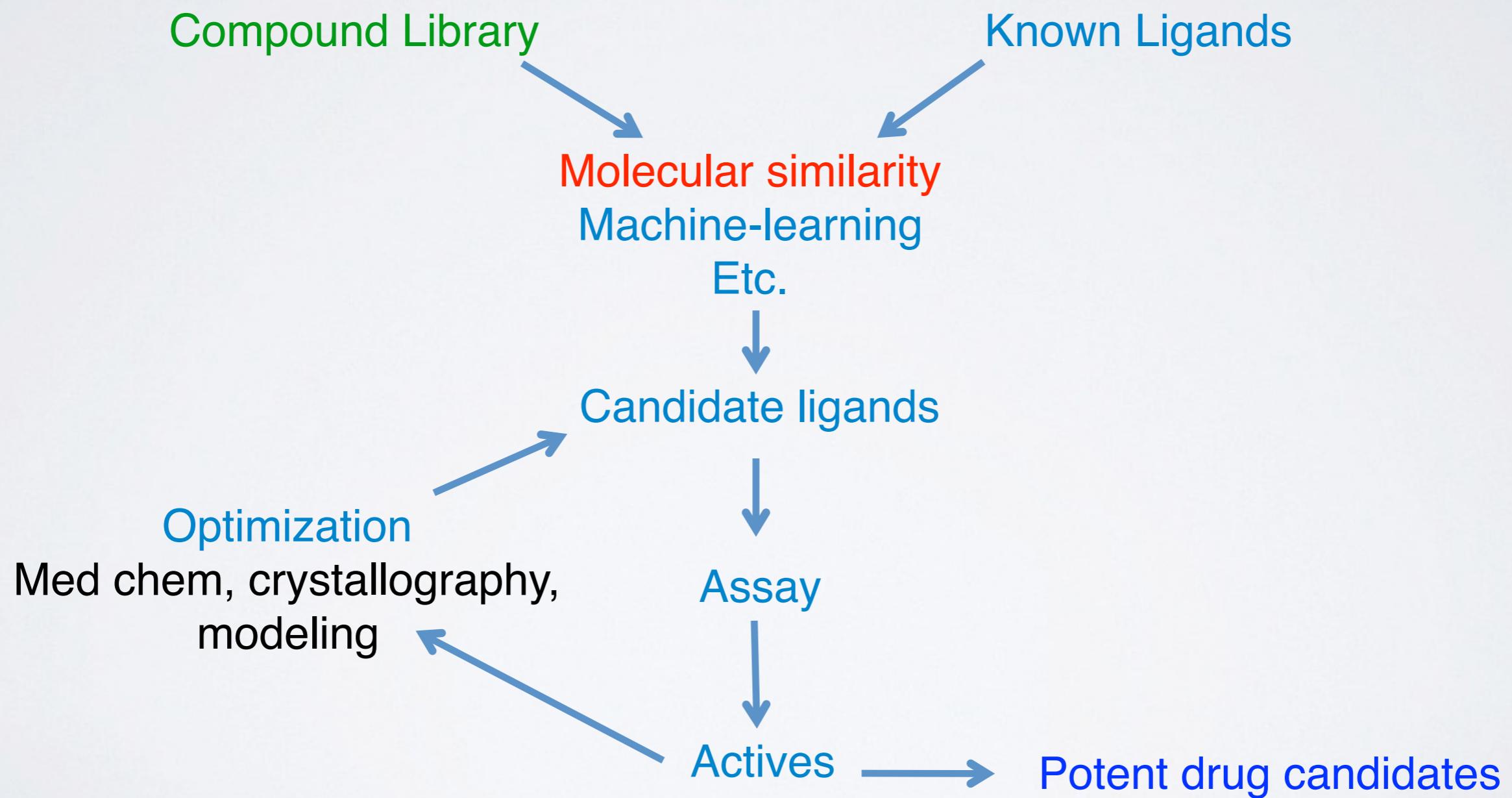
Why Look for Another Ligand if You Already Have Some?

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

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

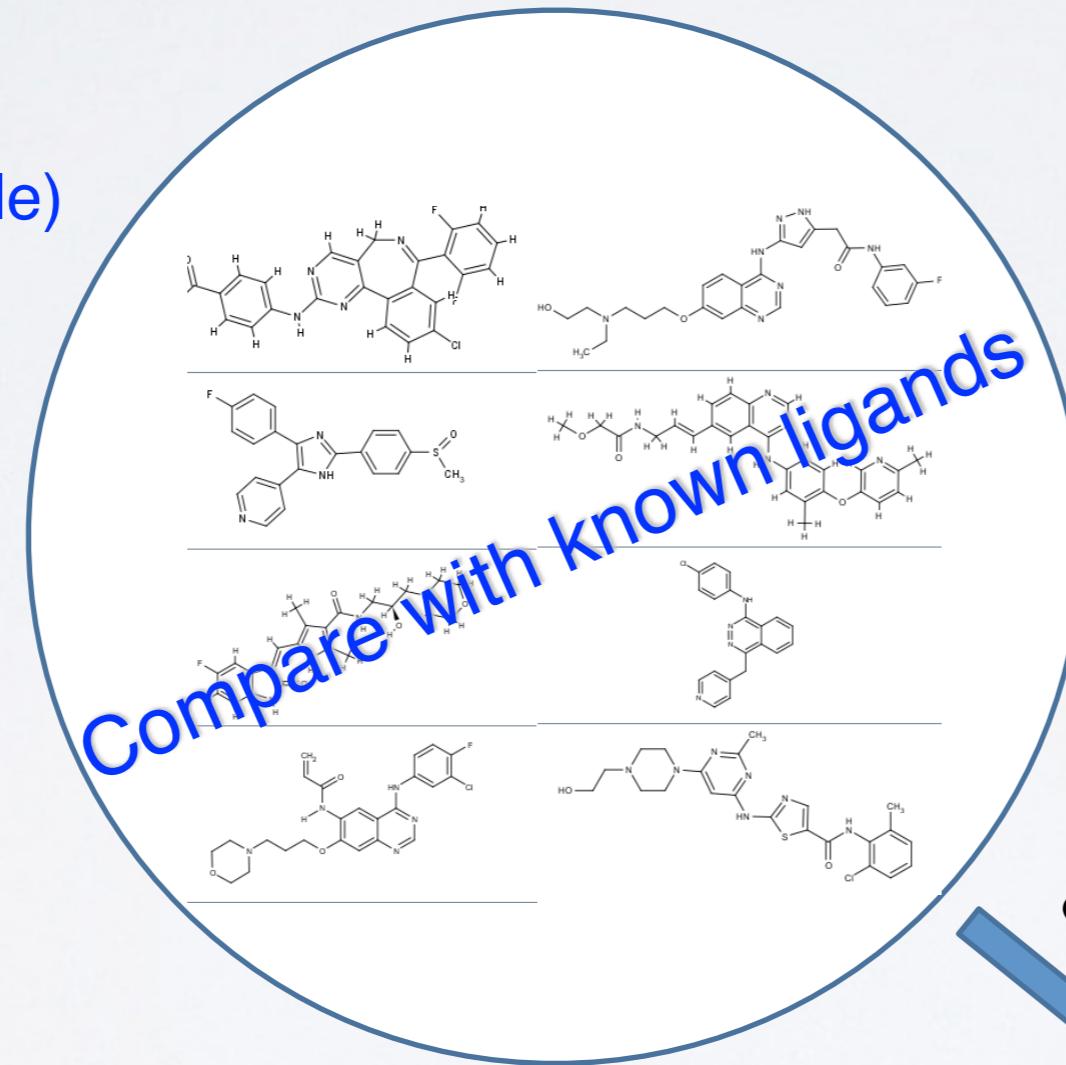
An high-affinity ligand is toxic, is not well-absorbed, difficult to synthesize etc.

LIGAND-BASED VIRTUAL SCREENING



CHEMICAL SIMILARITY LIGAND-BASED DRUG-DISCOVERY

Compounds
(available/synthesizable)



Different

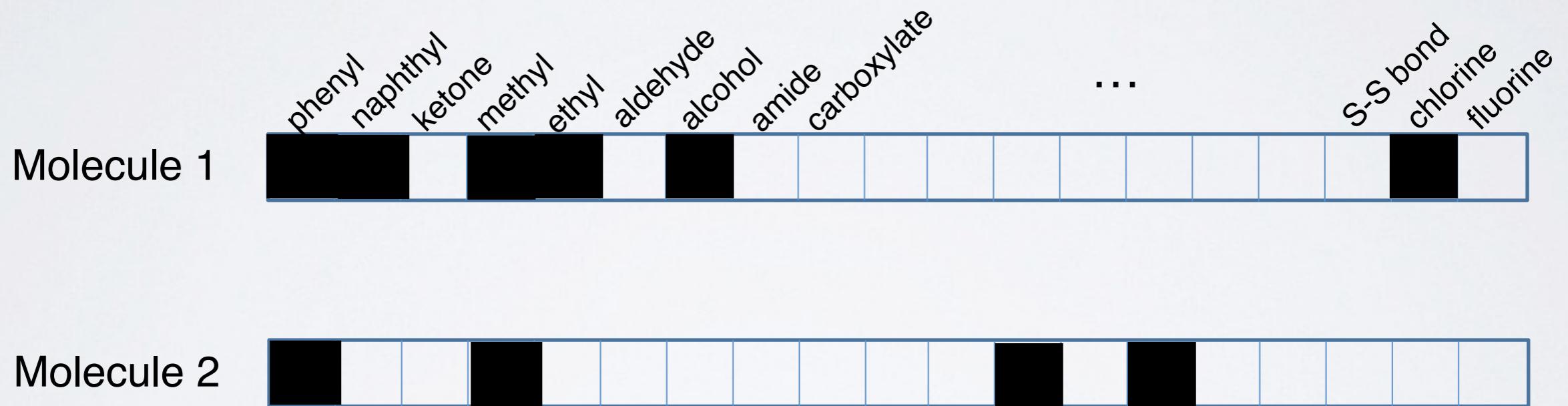
Don't bother

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



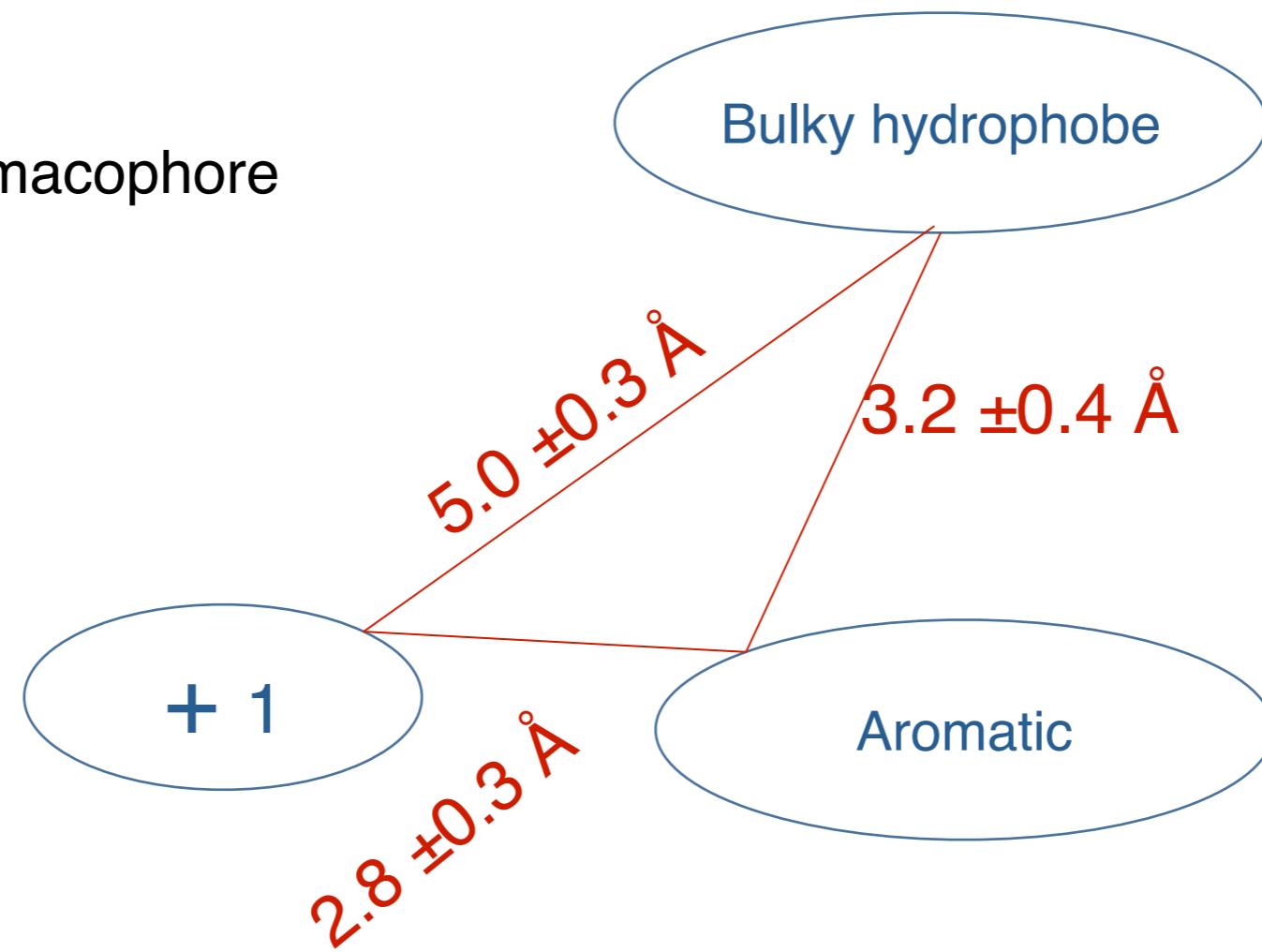
Union



Pharmacophore Models

Φάρμακο (drug) + Φορά (carry)

A 3-point pharmacophore



Molecular Descriptors

More abstract than chemical fingerprints

Physical descriptors

molecular weight

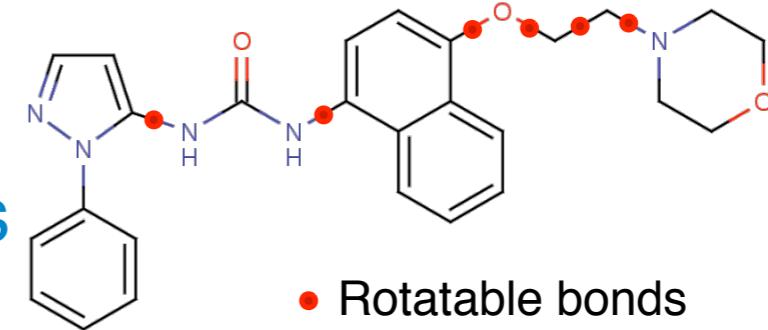
charge

dipole moment

number of H-bond donors/acceptors

number of rotatable bonds

hydrophobicity (log P and clogP)



- Rotatable bonds

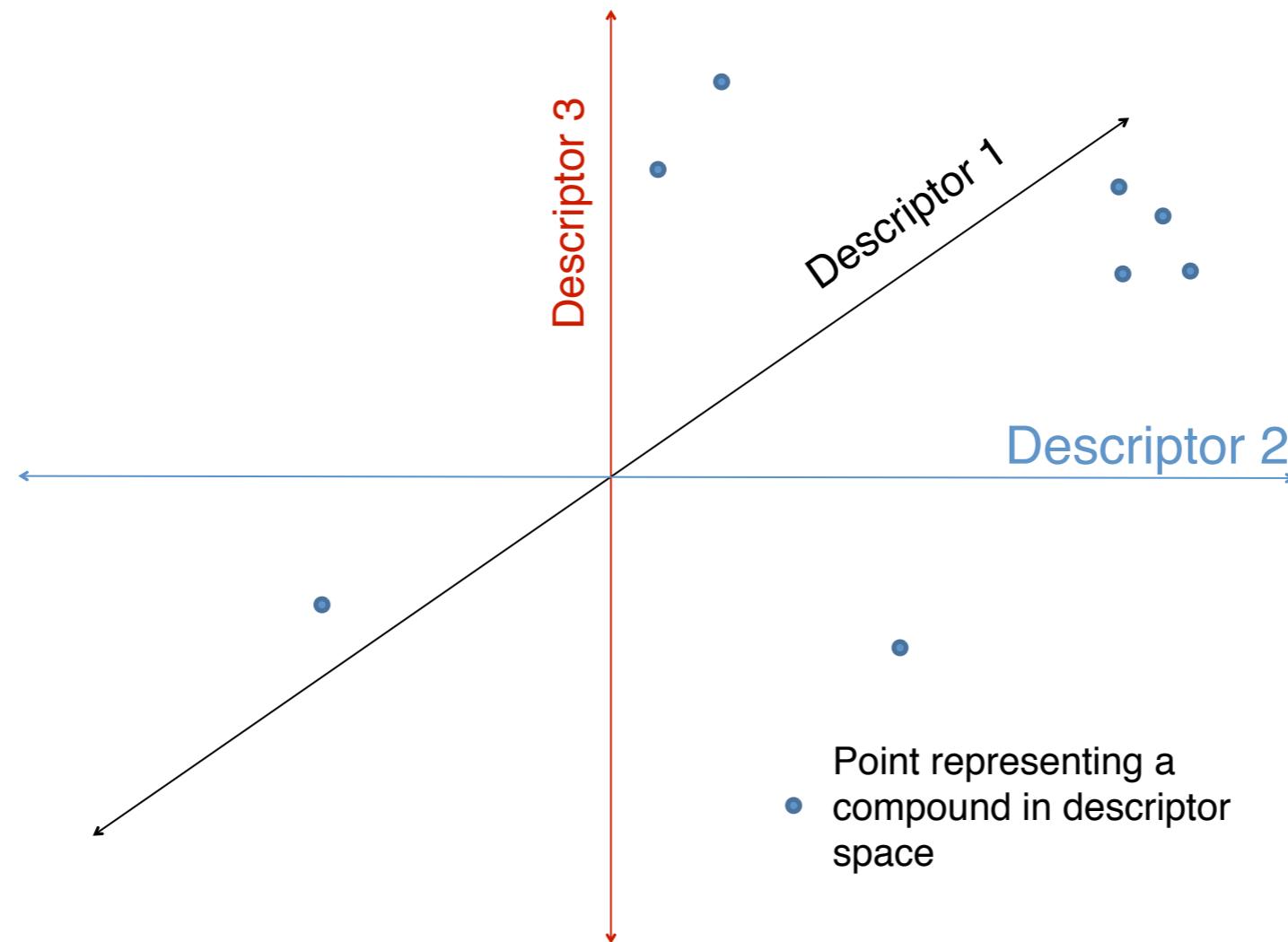
Topological
branching index
measures of linearity vs interconnectedness

Etc. etc.

A High-Dimensional “Chemical Space”

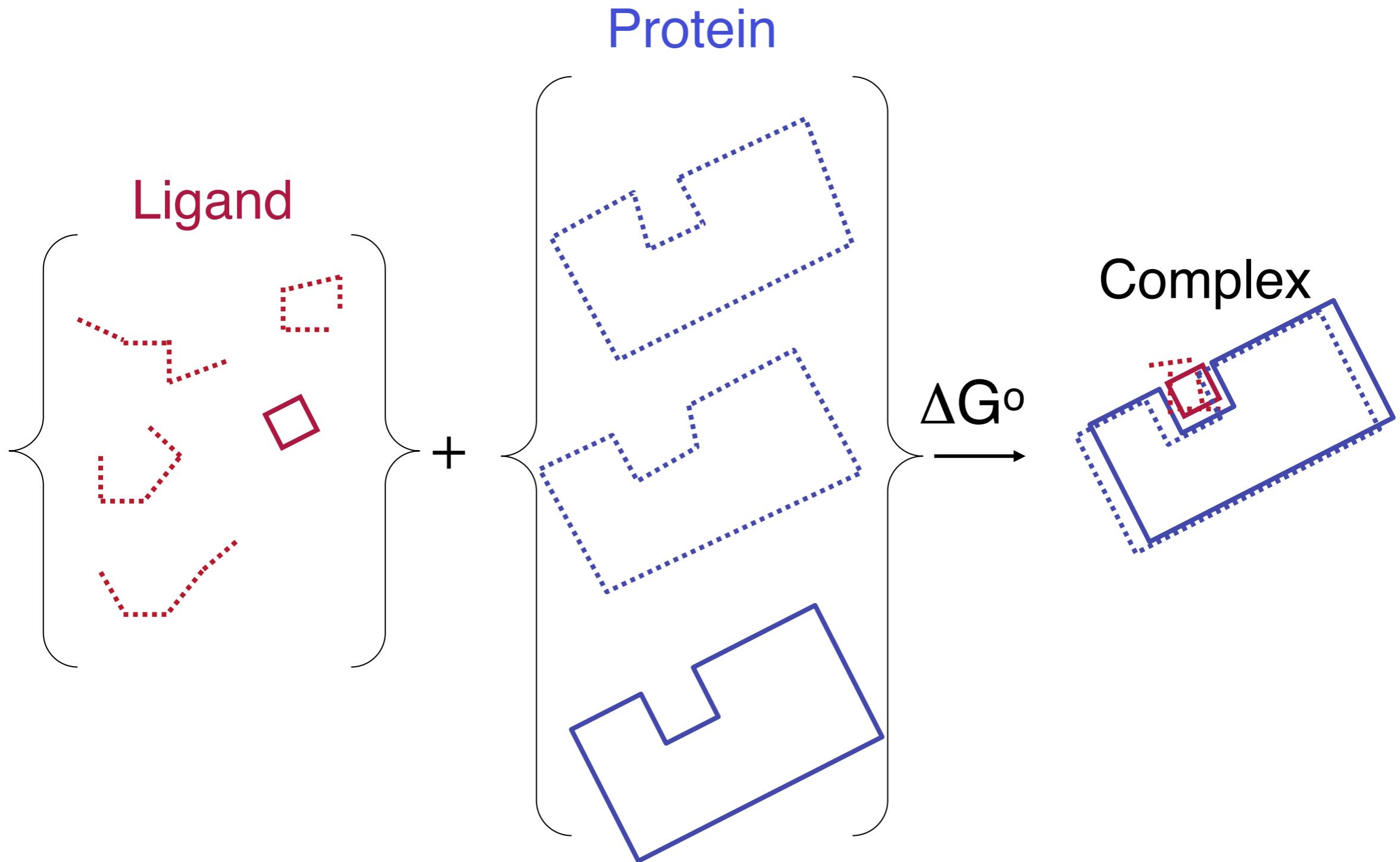
Each compound is a point in an n-dimensional space

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



Do it Yourself!

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



https://bioboot.github.io/bggm213_S19/lectures/#12

NMA in Bio3D

- Normal Mode Analysis (NMA) is a bioinformatics method that can predict the major motions of biomolecules.

```
pdb <- read.pdb("1hel")
modes <- nma( pdb )
m7 <- mktrj(modes, mode=7, file="mode_7.pdb")
```

Then you can open the resulting **mode_7.pdb** file in **VMD**
- Use "TUBE" representation and hit the play button...

Or use the **bio3d.view view()** function

```
library("bio3d.view")
view(m7, col=vec2color(rmsf(m7)))
```

Reference Slides

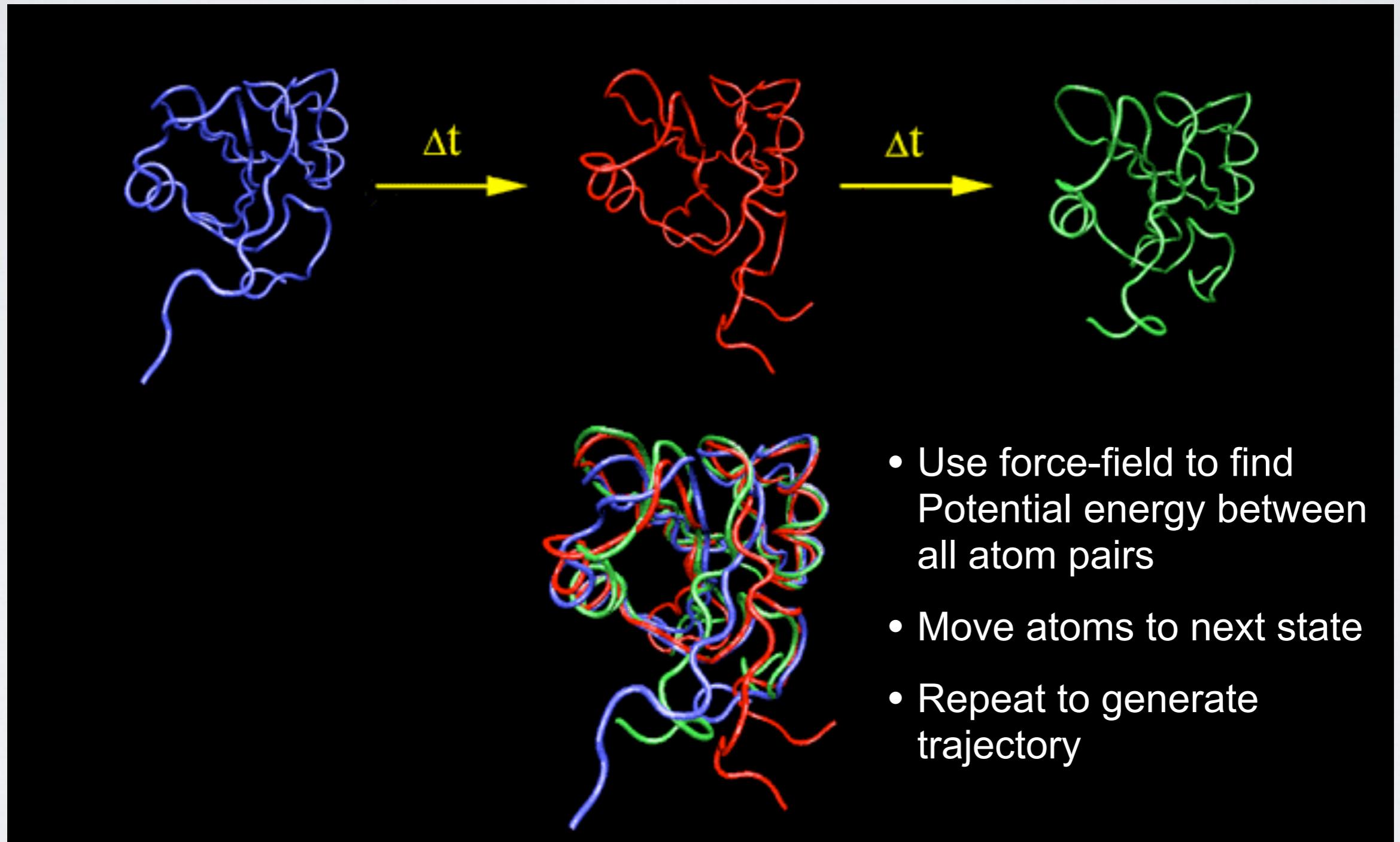
Molecular Dynamics (MD) and Normal Mode Analysis
(NMA) Background and Cautionary Notes

[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

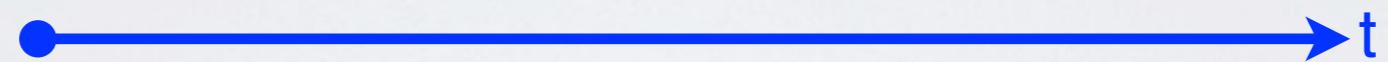
MOLECULAR DYNAMICS SIMULATION



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

[See: <https://www.youtube.com/watch?v=ui1ZysMFcKk>]

- ▶ Divide **time** into discrete ($\sim 1\text{fs}$) **time steps** (Δt)
(for integrating equations of motion, see below)



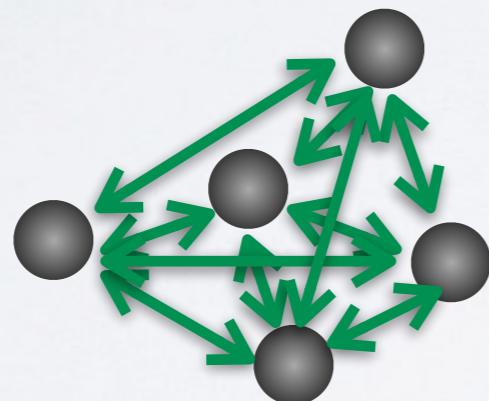
- ▶ Divide **time** into discrete ($\sim 1\text{fs}$) **time steps** (Δt)
(for integrating equations of motion, see below)



- ▶ Divide **time** into discrete ($\sim 1\text{fs}$) **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

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

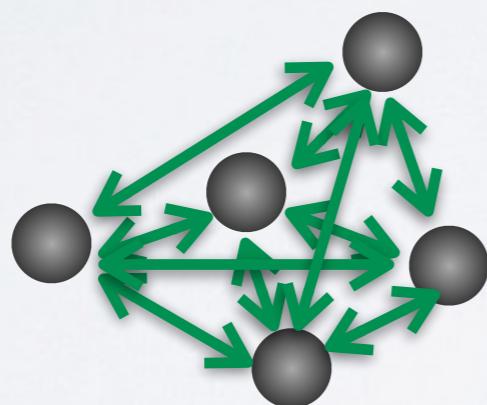
Empirical force field

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

- ▶ Divide time into discrete ($\sim 1\text{fs}$) 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)



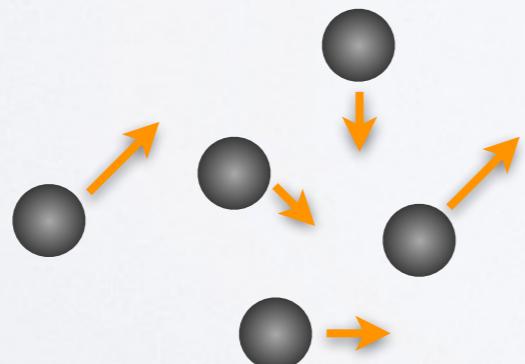
Nucleic motion described classically

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

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)



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

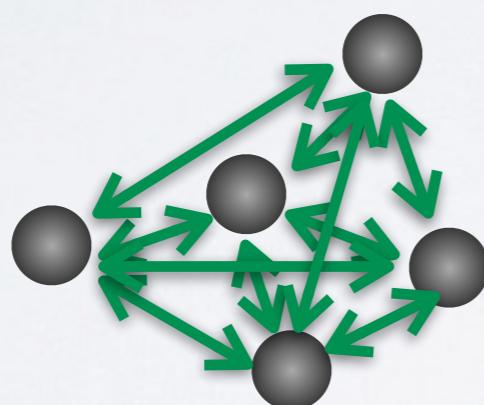
$$r(t + \Delta t) = r(t) + \boxed{v(t + \frac{\Delta t}{2})} \Delta t$$

BASIC ANATOMY OF A MD SIMULATION

- ▶ Divide **time** into discrete ($\sim 1\text{fs}$) **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)



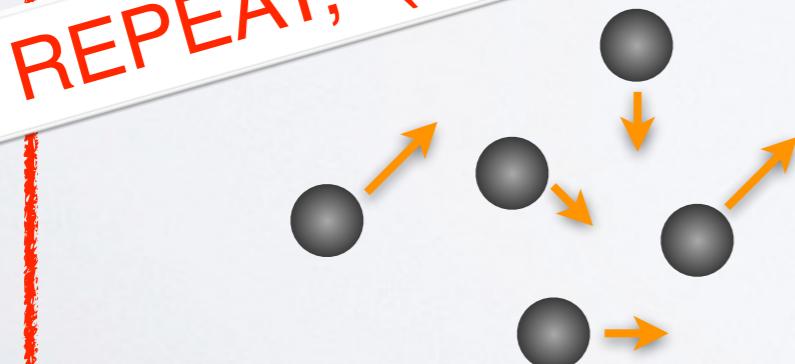
Nucleic motion described classically

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

Empirical force function

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

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

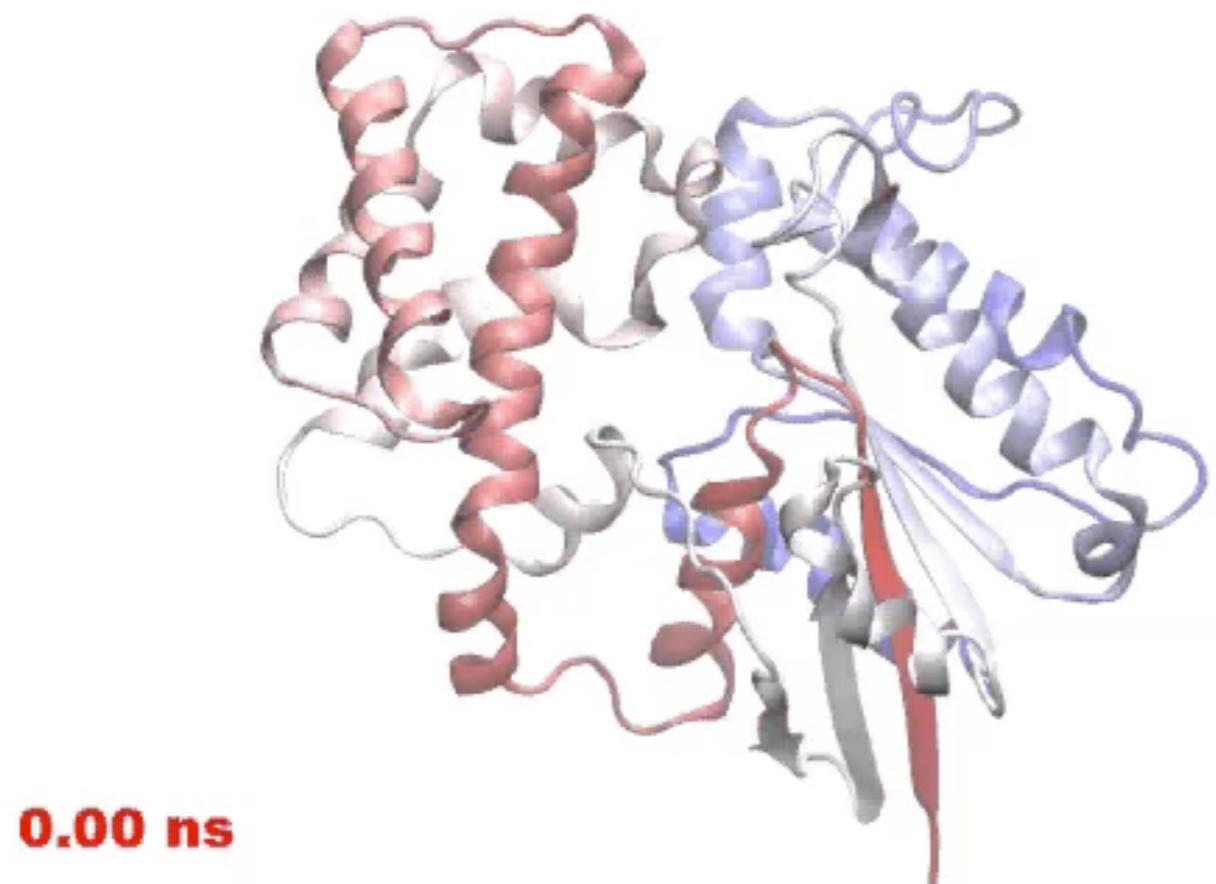


$$\begin{aligned} \mathbf{v}(t + \frac{\Delta t}{2}) &= \mathbf{v}(t - \frac{\Delta t}{2}) + \frac{\mathbf{F}(t)}{m} \Delta t \\ \mathbf{r}(t + \Delta t) &= \mathbf{r}(t) + \mathbf{v}(t + \frac{\Delta t}{2}) \Delta t \end{aligned}$$

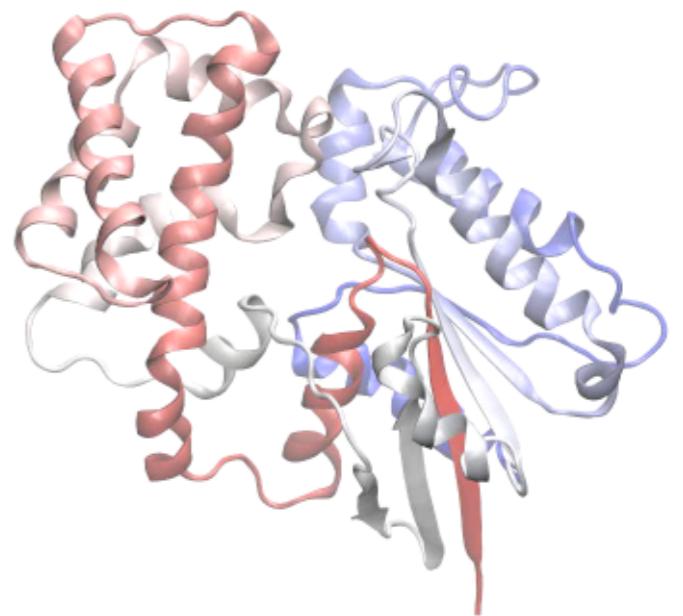
REPEAT, (iterate many, many times... 1ms = 10^{12} time steps)

MD Prediction of Functional Motions

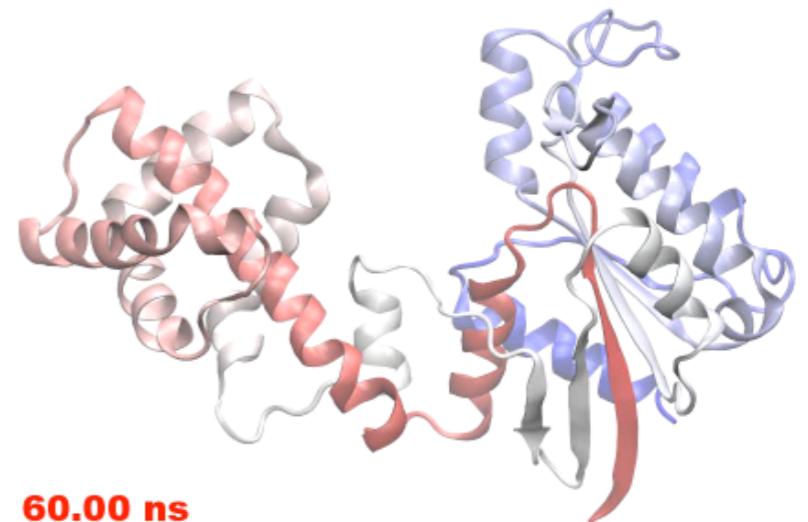
Accelerated MD simulation of
nucleotide-free transducin alpha subunit



“close”

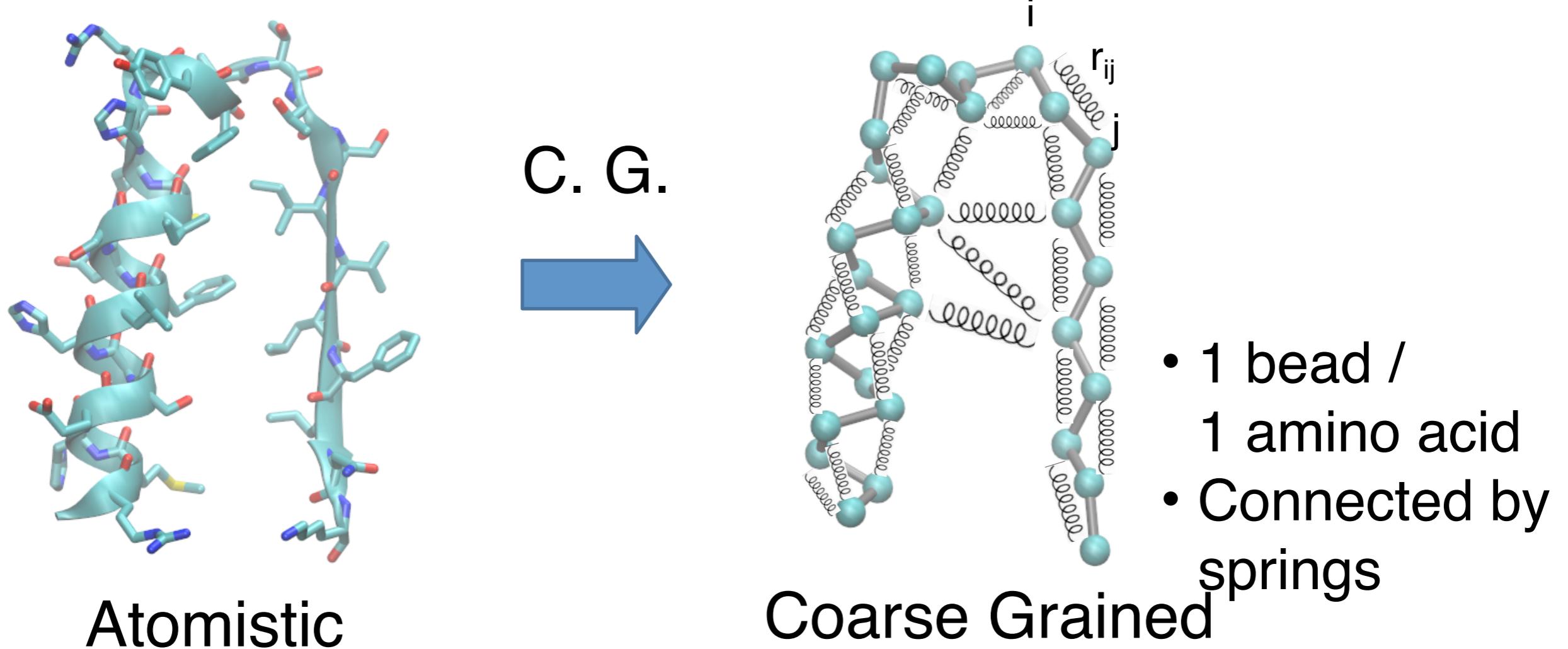


“open”

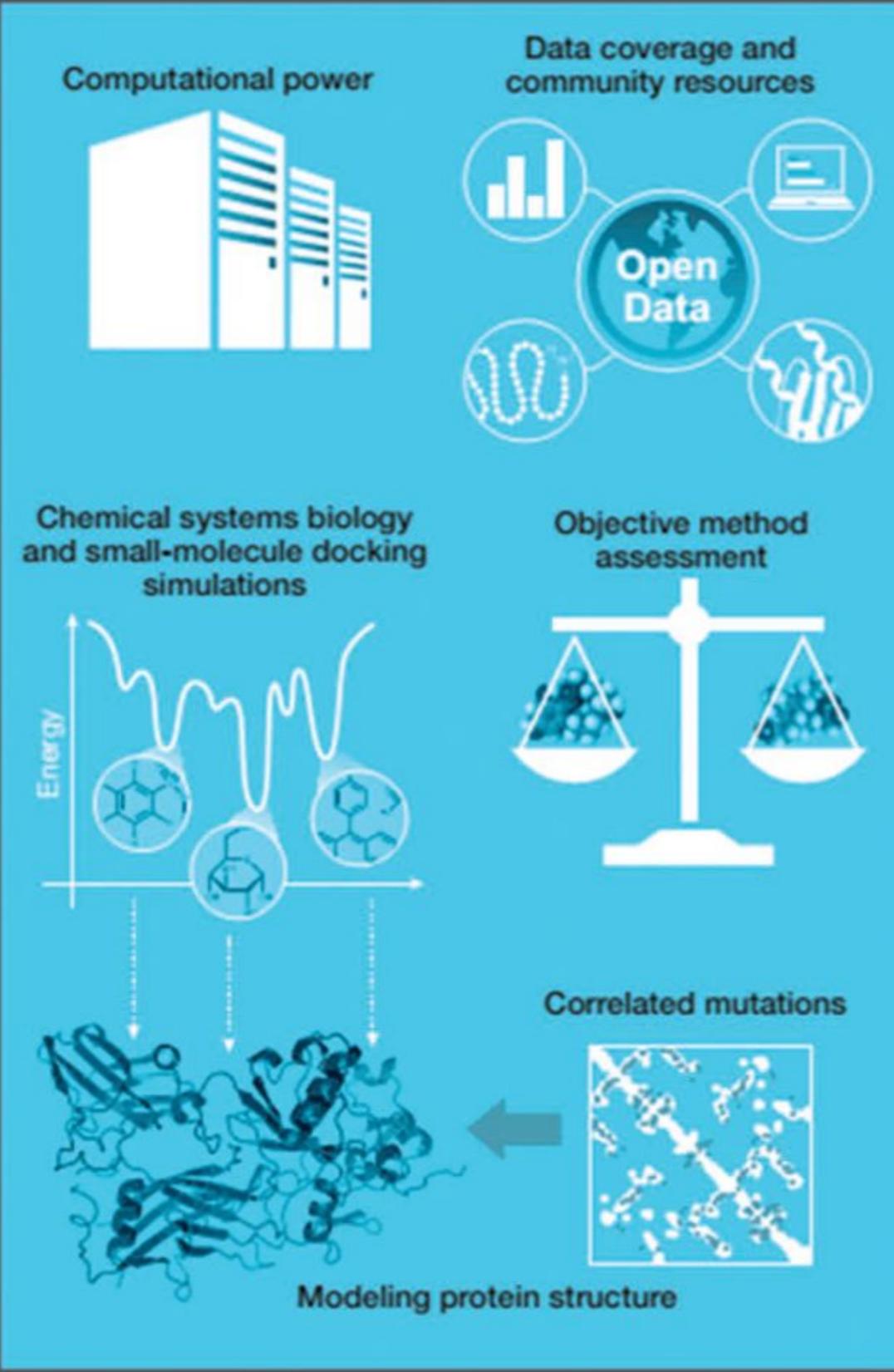


COARSE GRAINING: **NORMAL MODE ANALYSIS** (NMA)

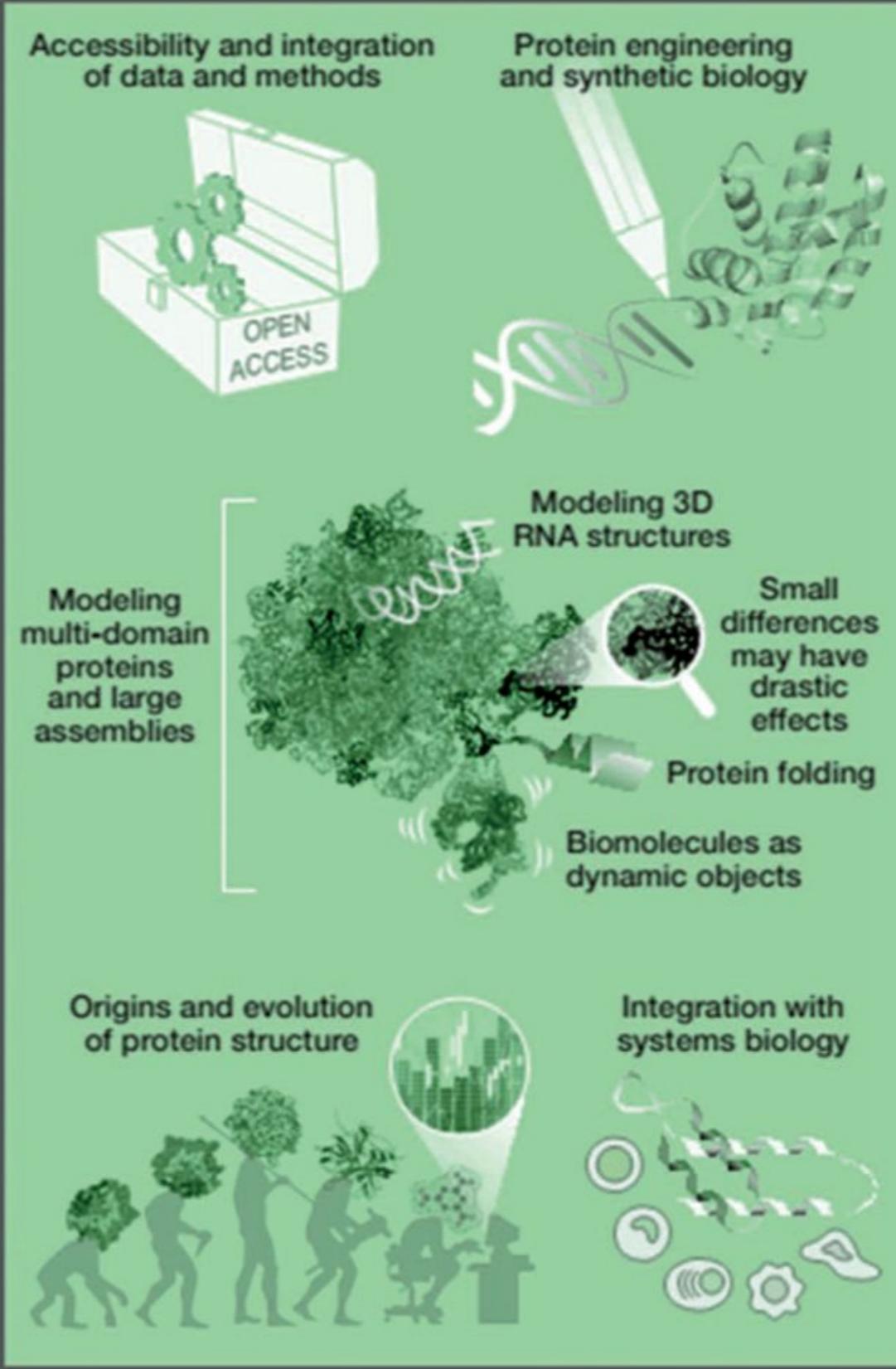
- MD is still time-consuming for large systems
- Elastic network model NMA (ENM-NMA) is an example of a lower resolution approach that finishes in seconds even for large systems.



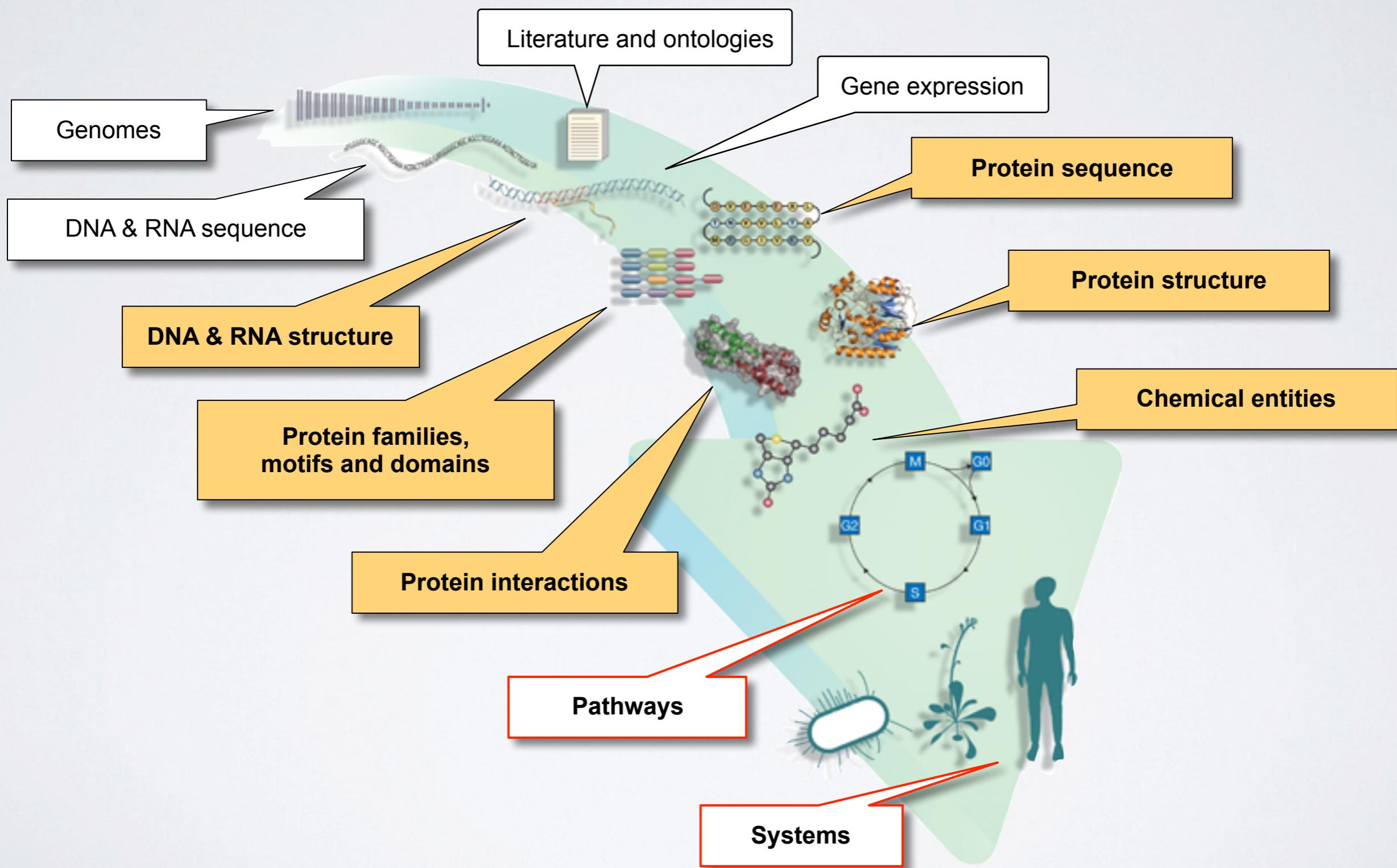
ACHIEVEMENTS



CHALLENGES



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.