BGGN 213

Structural Bioinformatics II

Lecture 13

Barry Grant

UC San Diego

http://thegrantlab.org/bggn213

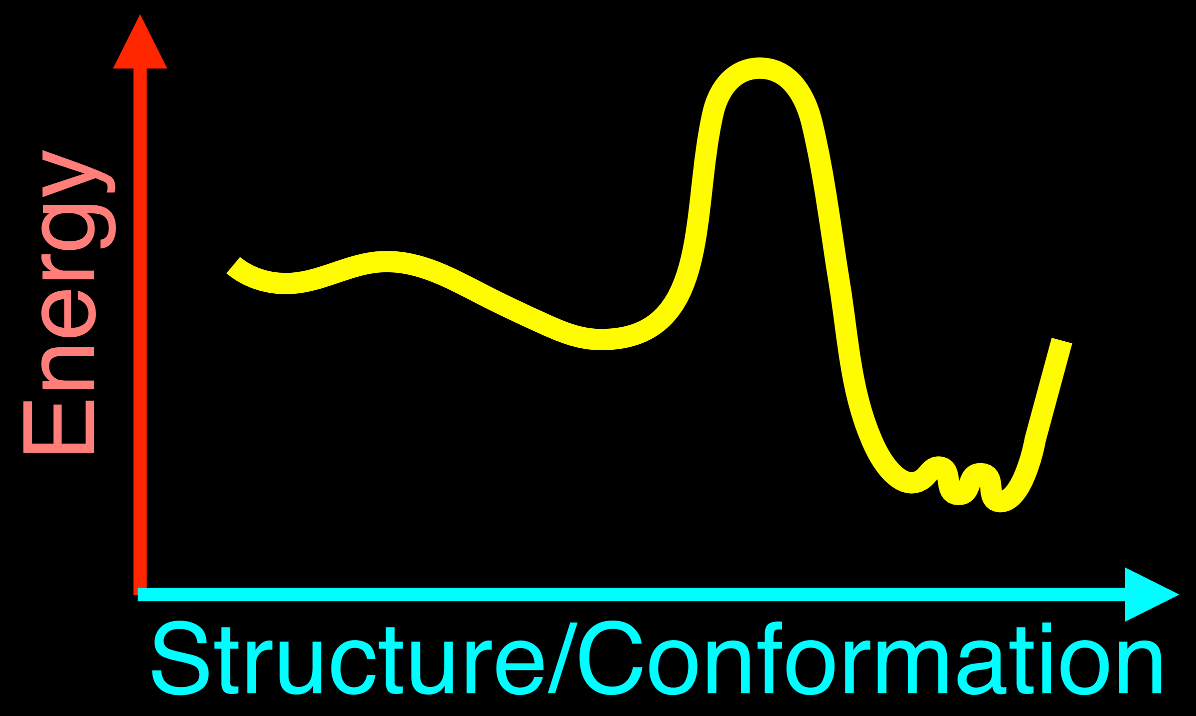
Download MGL Tools: See class website!

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:

$\boxed{\text{(1). Physics-Based}}$

(2). Knowledge-Based

For physics based potentials
energy terms come from physical theory

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

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

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

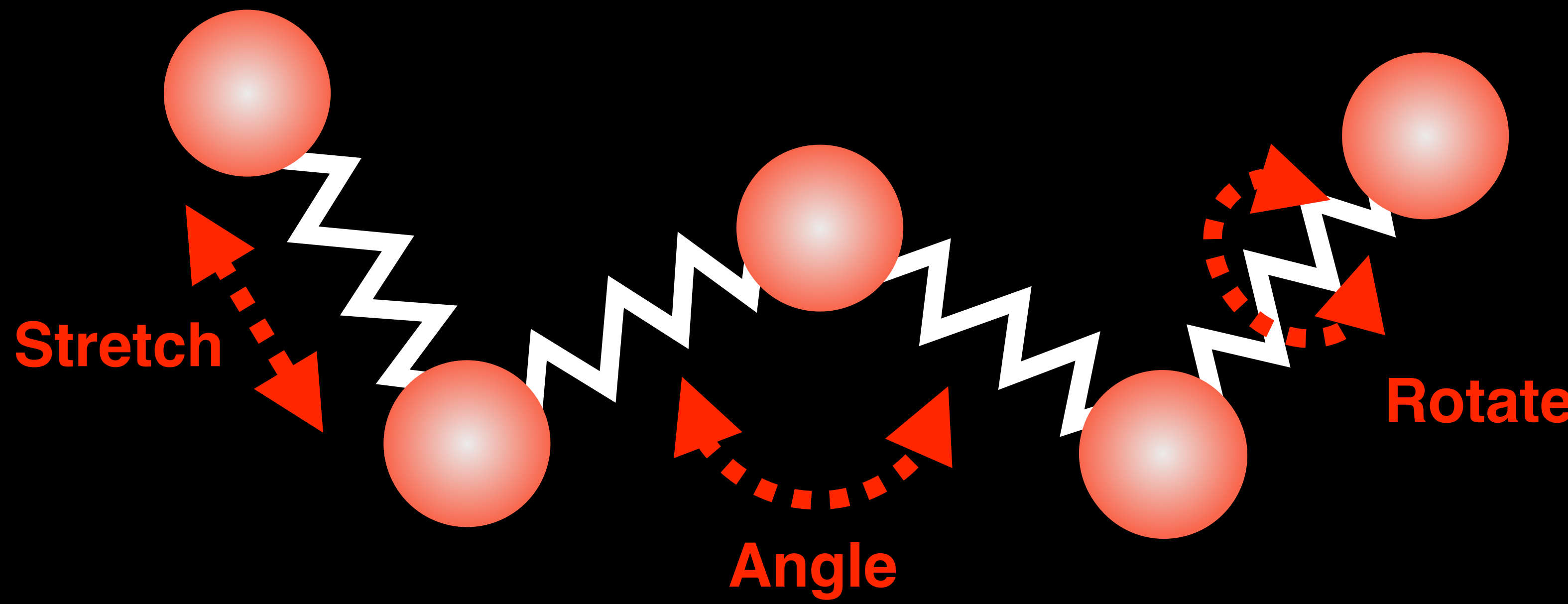
$$V(R) = \boxed{E_{bonded}} + E_{non.bonded}$$

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

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

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

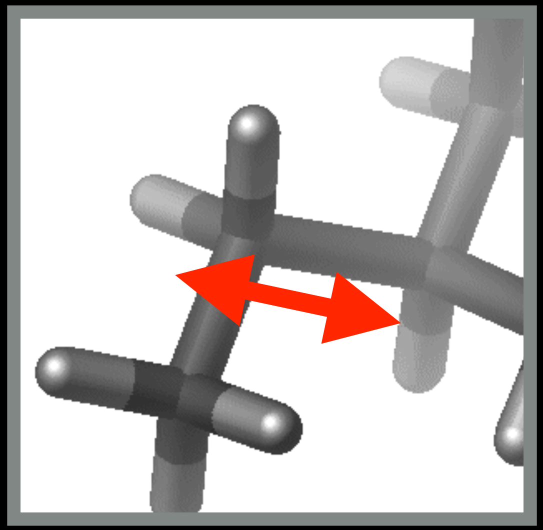
$$\boxed{E_{bond.stretch} + E_{bond.angle} + E_{bond.rotate}}$$



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

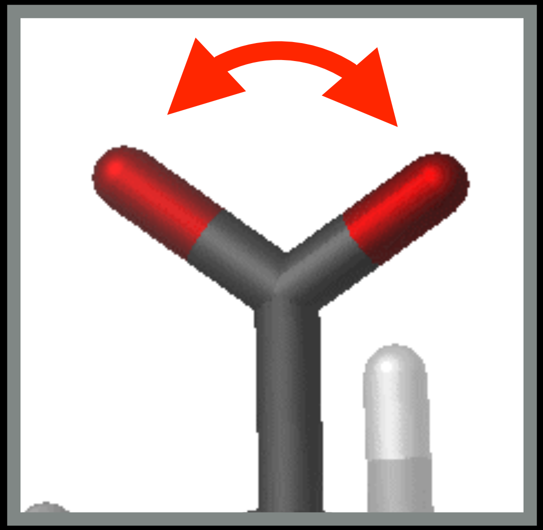
E_{bonded} **is itself a sum of three terms:**

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



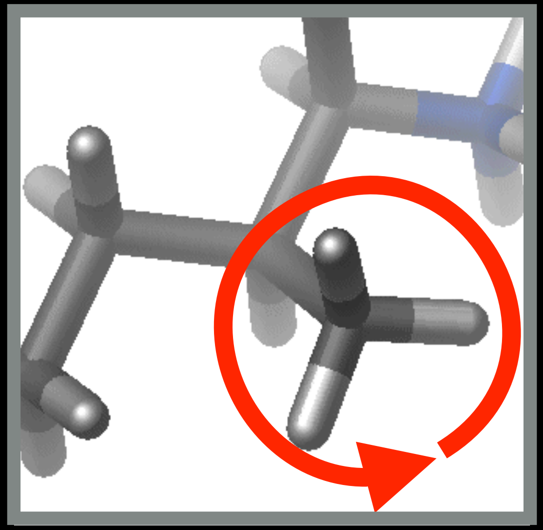
Bond Stretch

$$E_{bond.stretch}$$



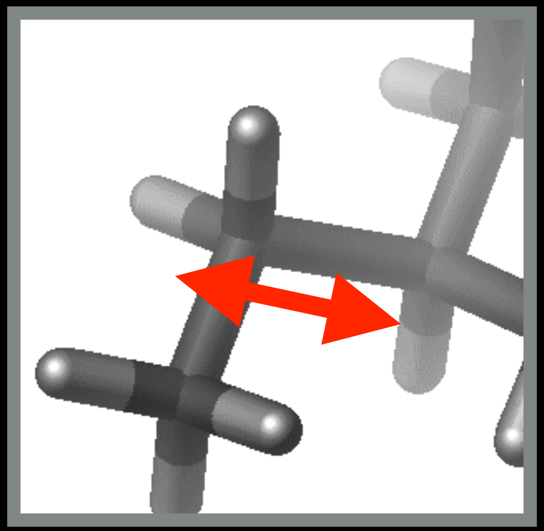
Bond Angle

$$E_{bond.angle}$$



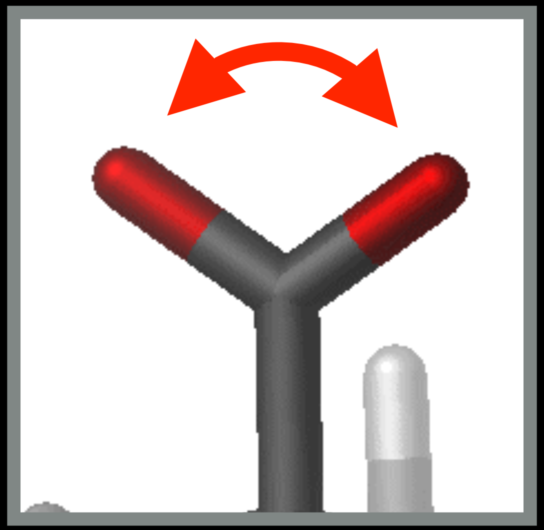
Bond Rotate

$$E_{bond.rotate}$$



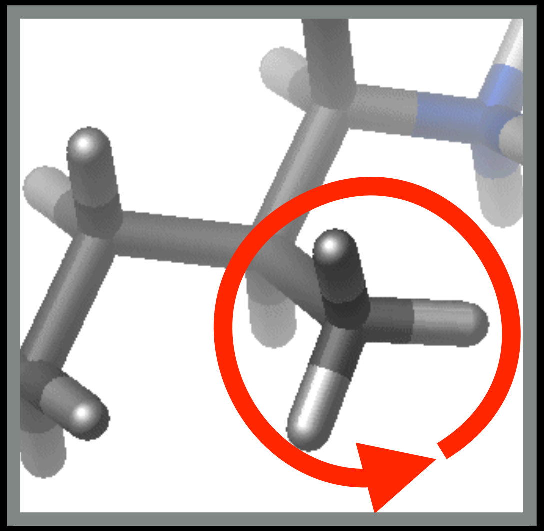
Bond Stretch

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



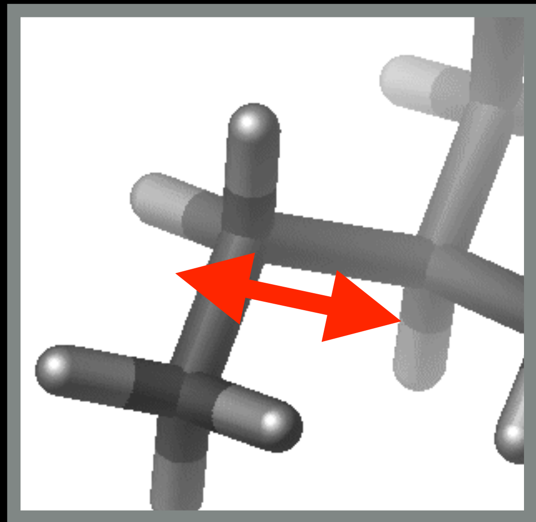
Bond Angle

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



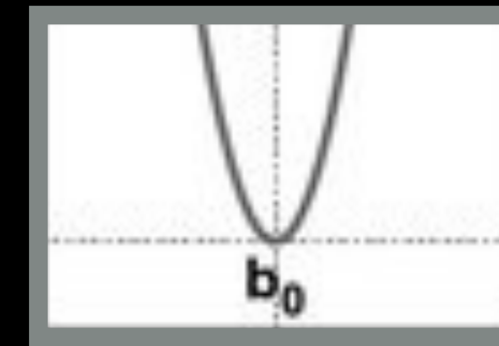
Bond Rotate

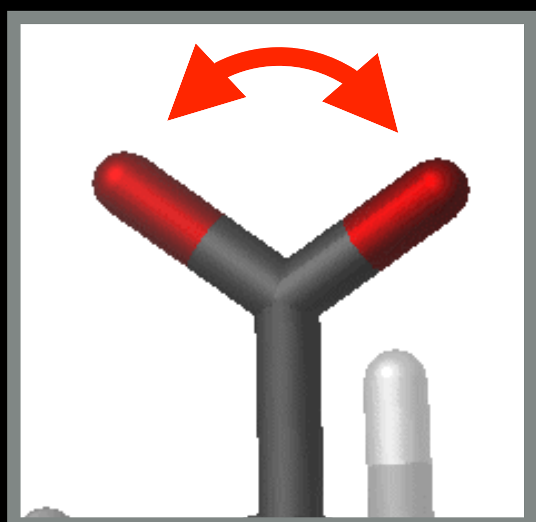
$$\sum_{dihedrals} K_i^{br}[1 - cos(n_i\phi_i - \phi_o)$$



Bond Stretch

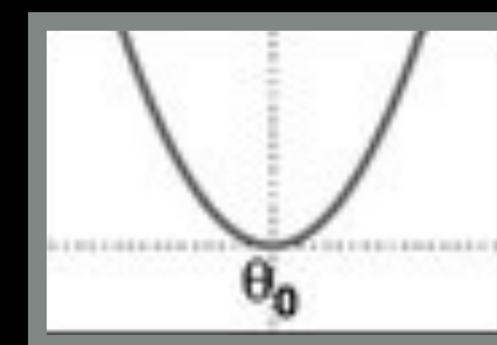
$$\sum_{bonds} K_i^{bs}(b_i - b_o)$$

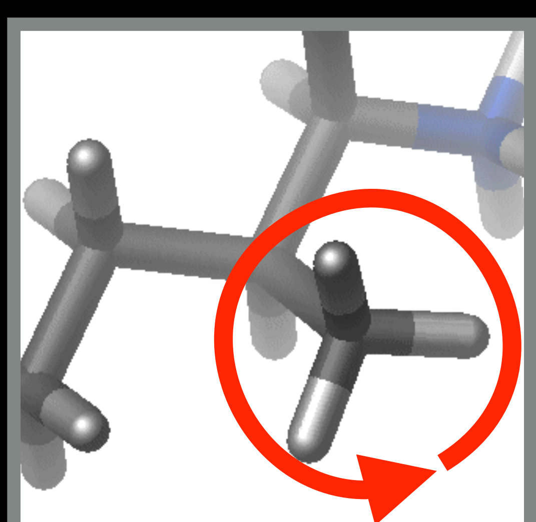




Bond Angle

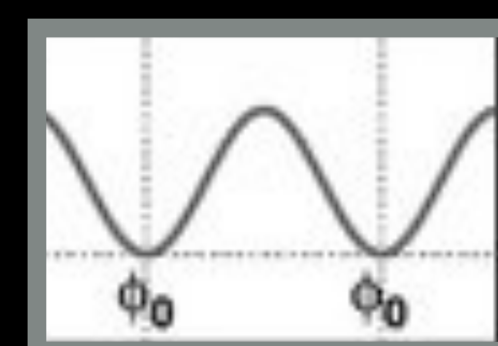
$$\sum_{angles} K_i^{ba}(\theta_i - \theta_o)$$





Bond Rotate

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



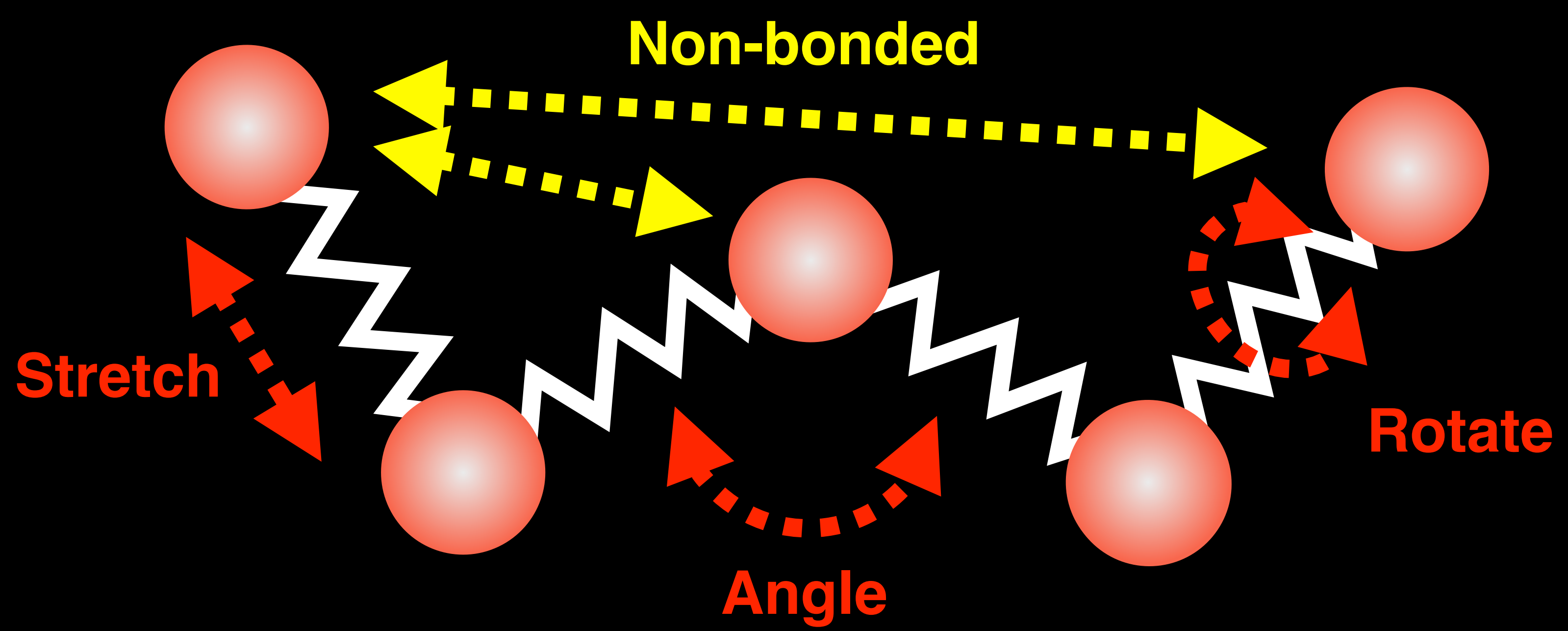
$$V(R) = E_{bonded} + \boxed{E_{non.bonded}}$$

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

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

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

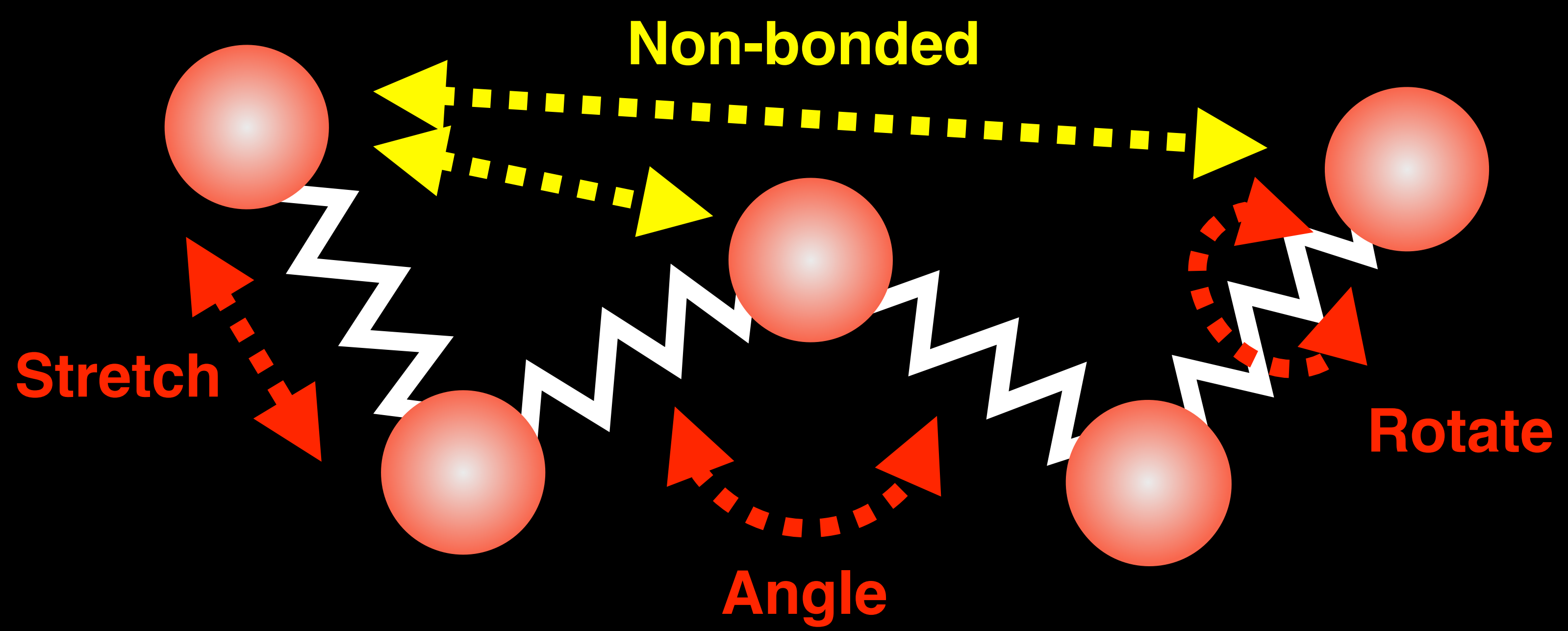
$$E_{van.der.Waals} + E_{electrostatic}$$



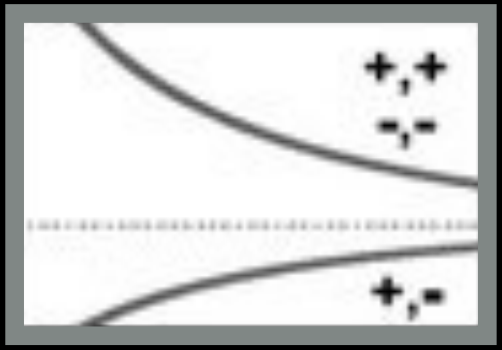
$$V(R) = E_{bonded} + \boxed{E_{non.bonded}}$$

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

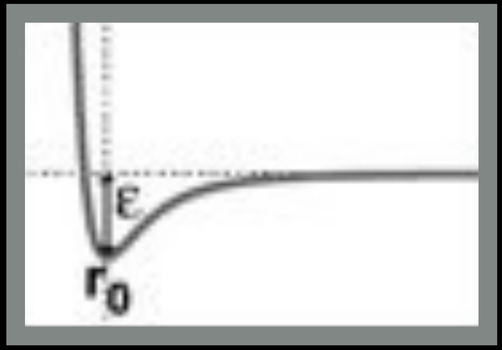
$$E_{van.der.Waals} + E_{electrostatic}$$



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



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



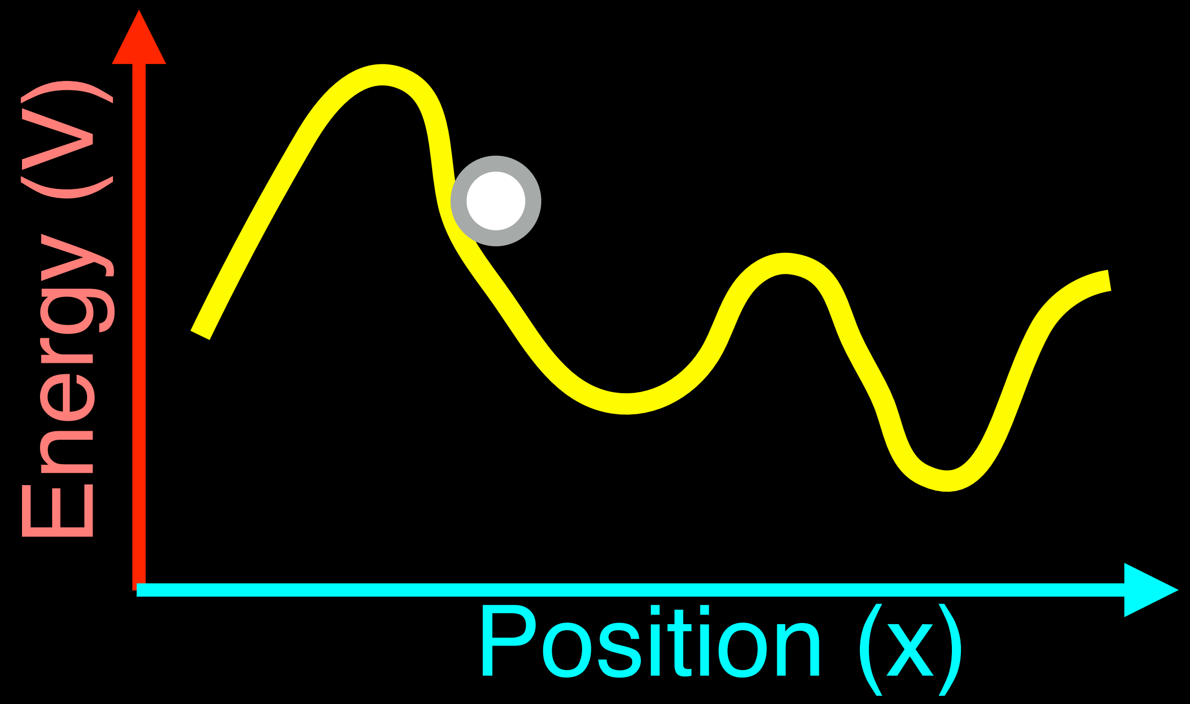
Total potential energy

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

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

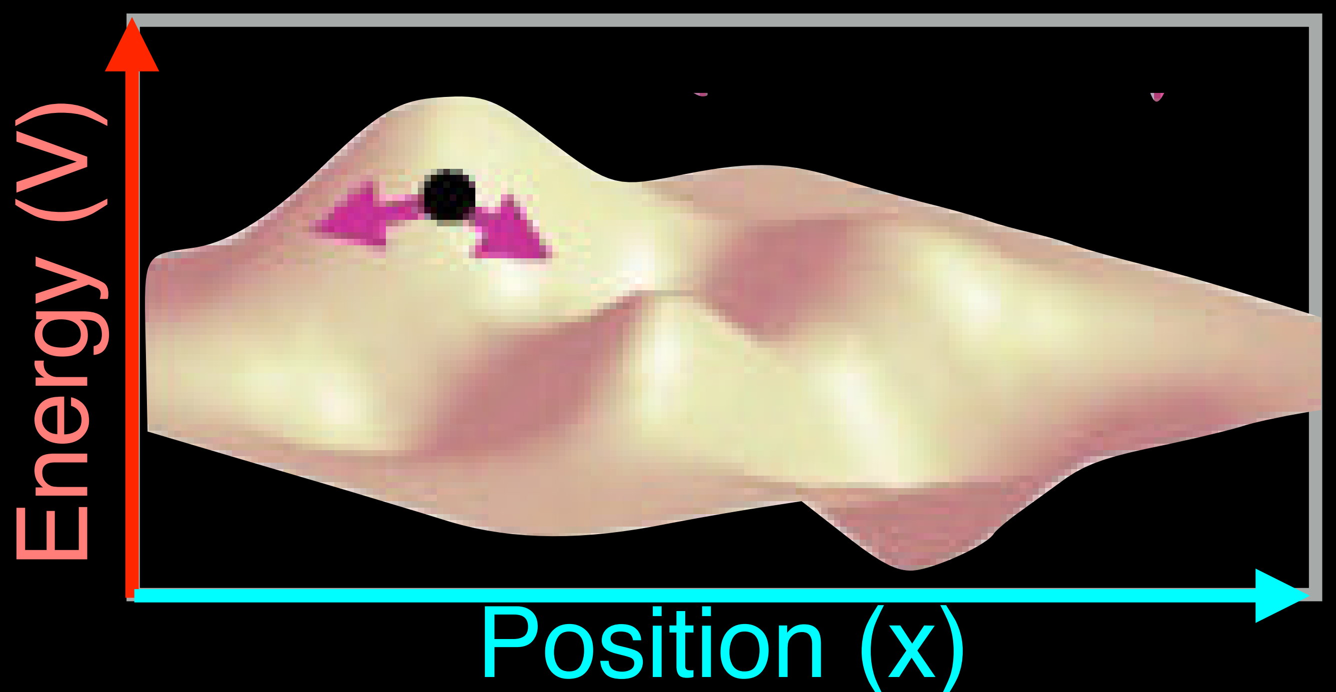
Potential energy surface

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



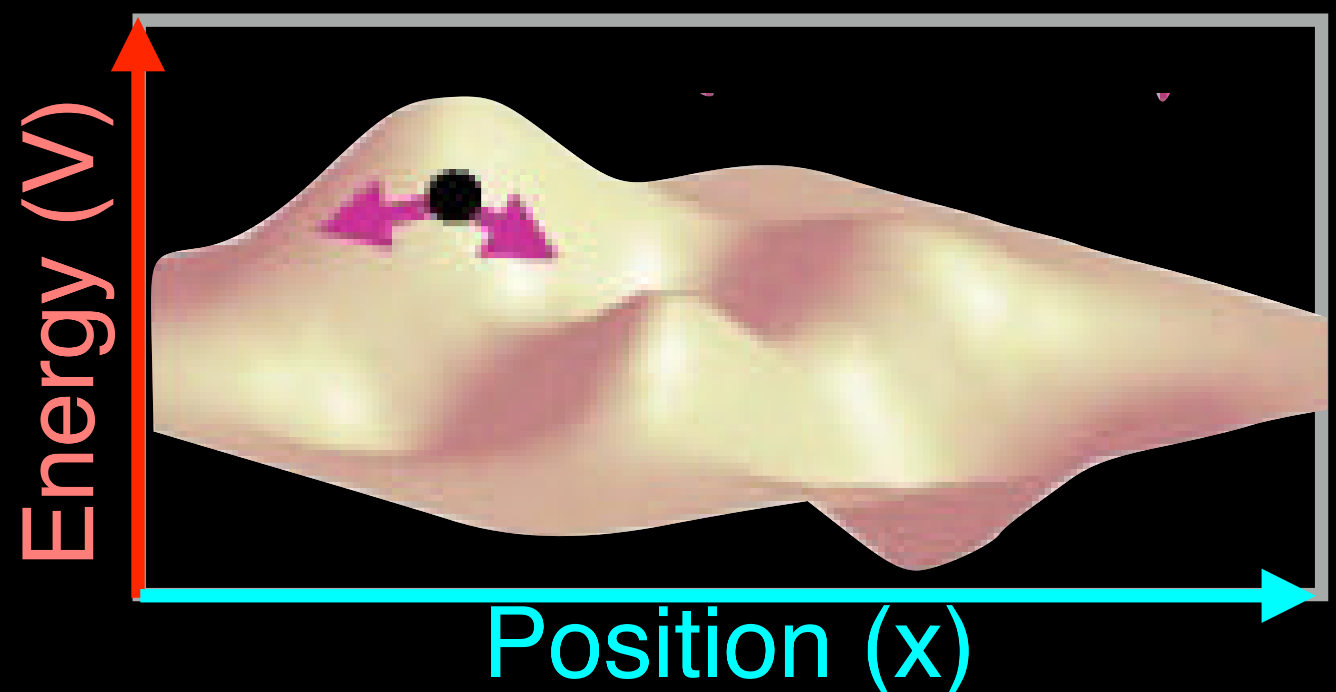
Potential energy surface

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

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

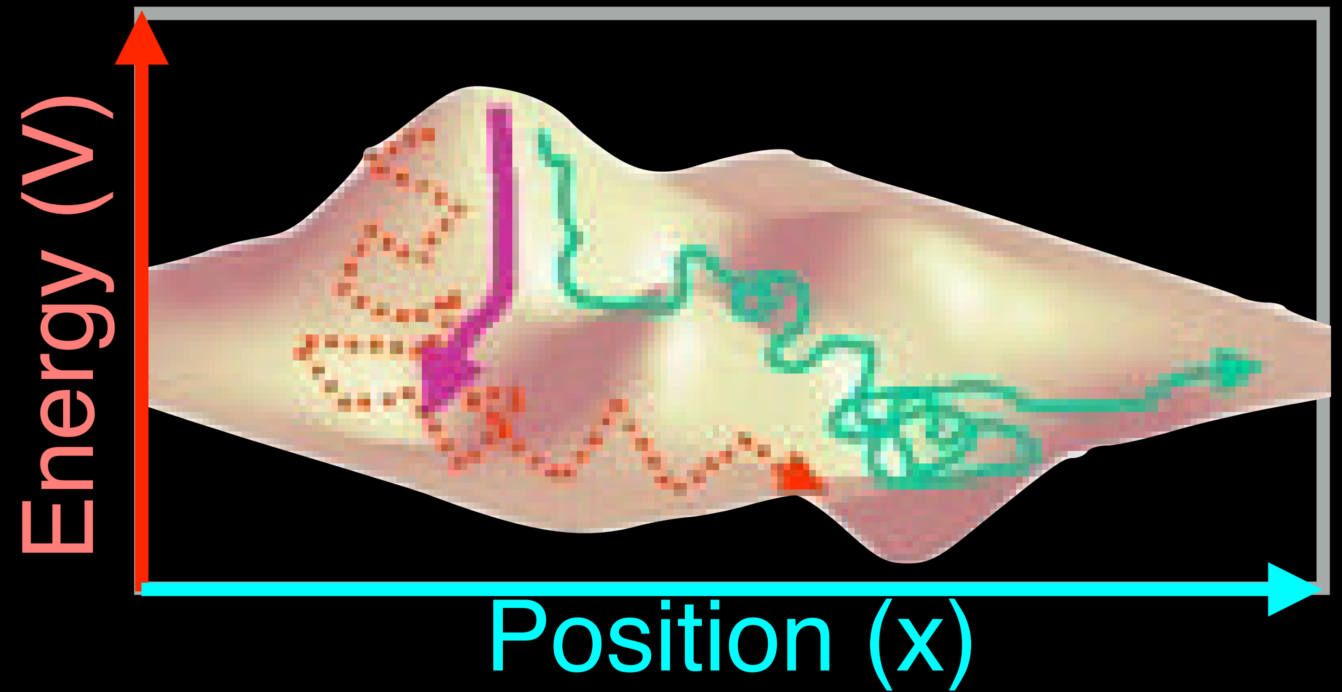
Moving Over The Energy Surface

- **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

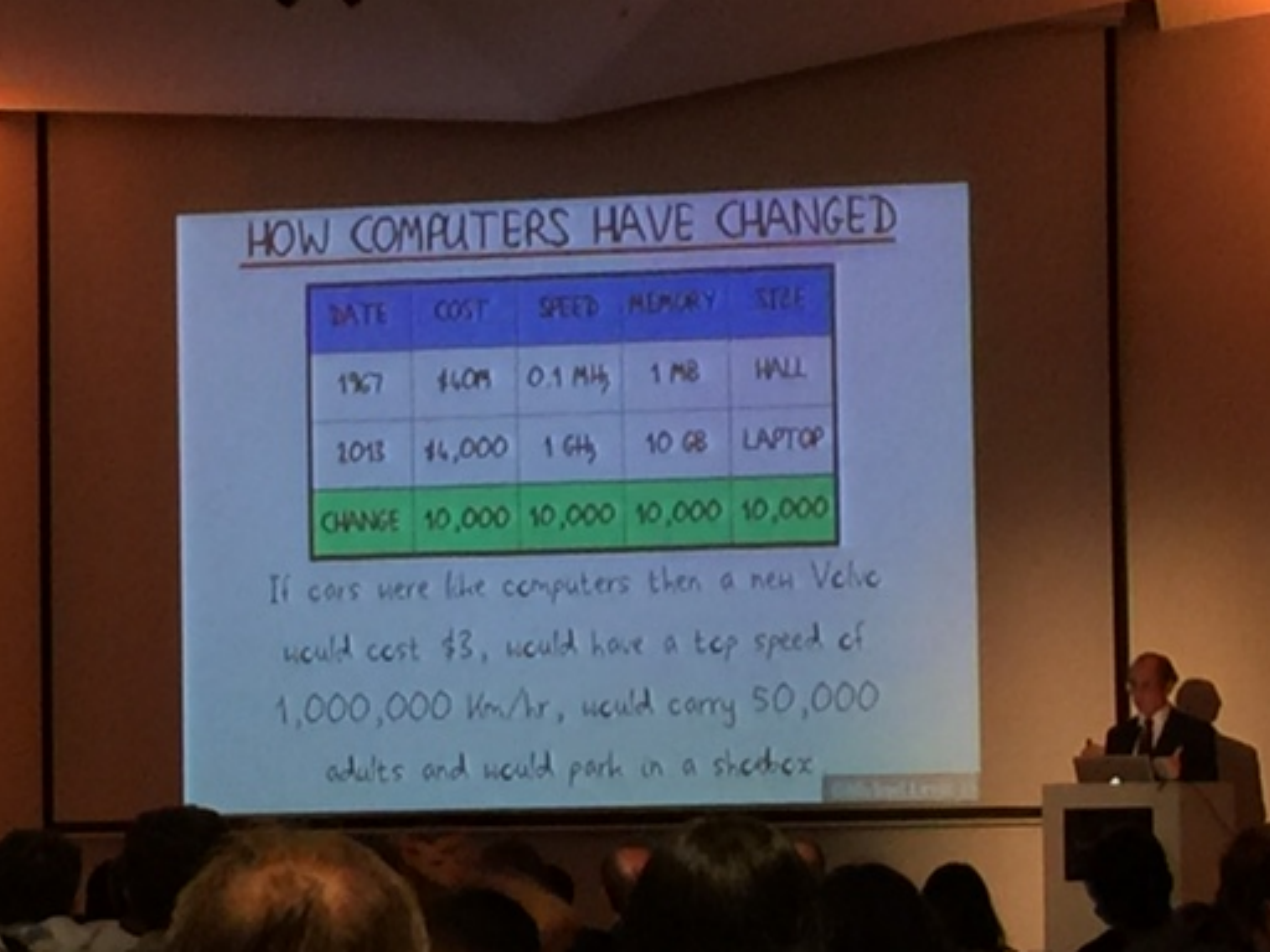
Fully physical detail becomes computationally intractable
Approximations are unavoidable
(Quantum effects approximated classically, water may be treated crudely)
Parameterization still required

Strengths

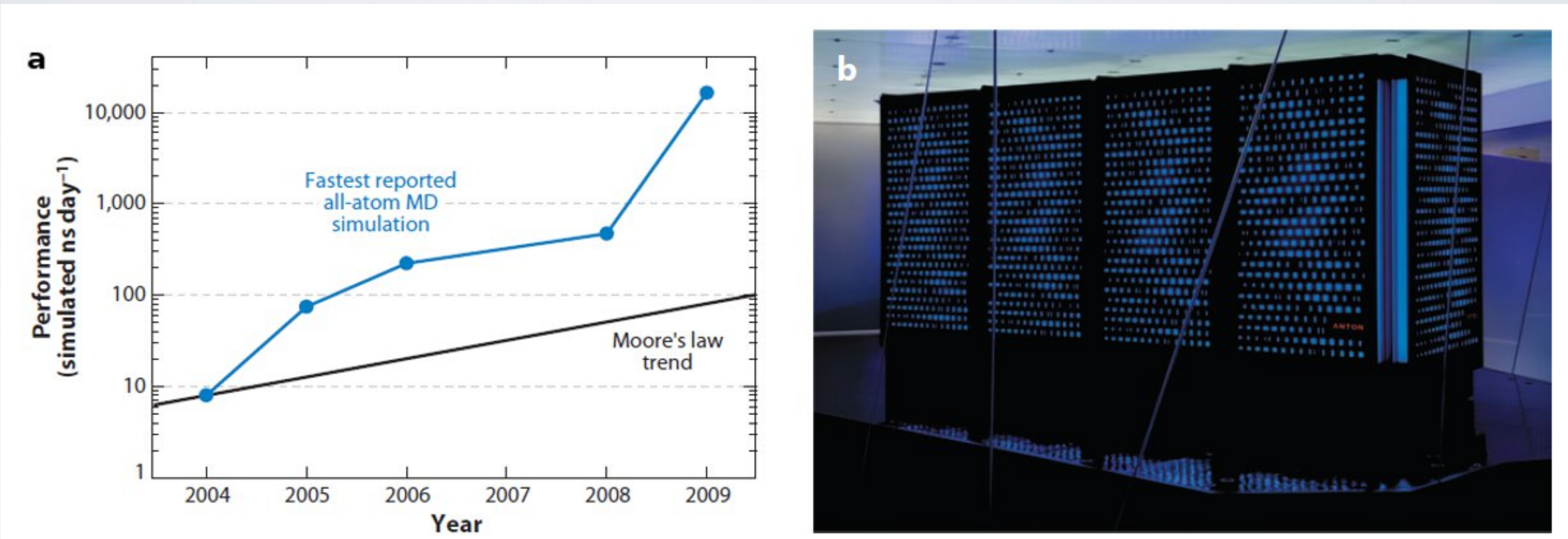
Interpretable, provides guides to design
Broadly applicable, in principle at least
Clear pathways to improving accuracy

Status

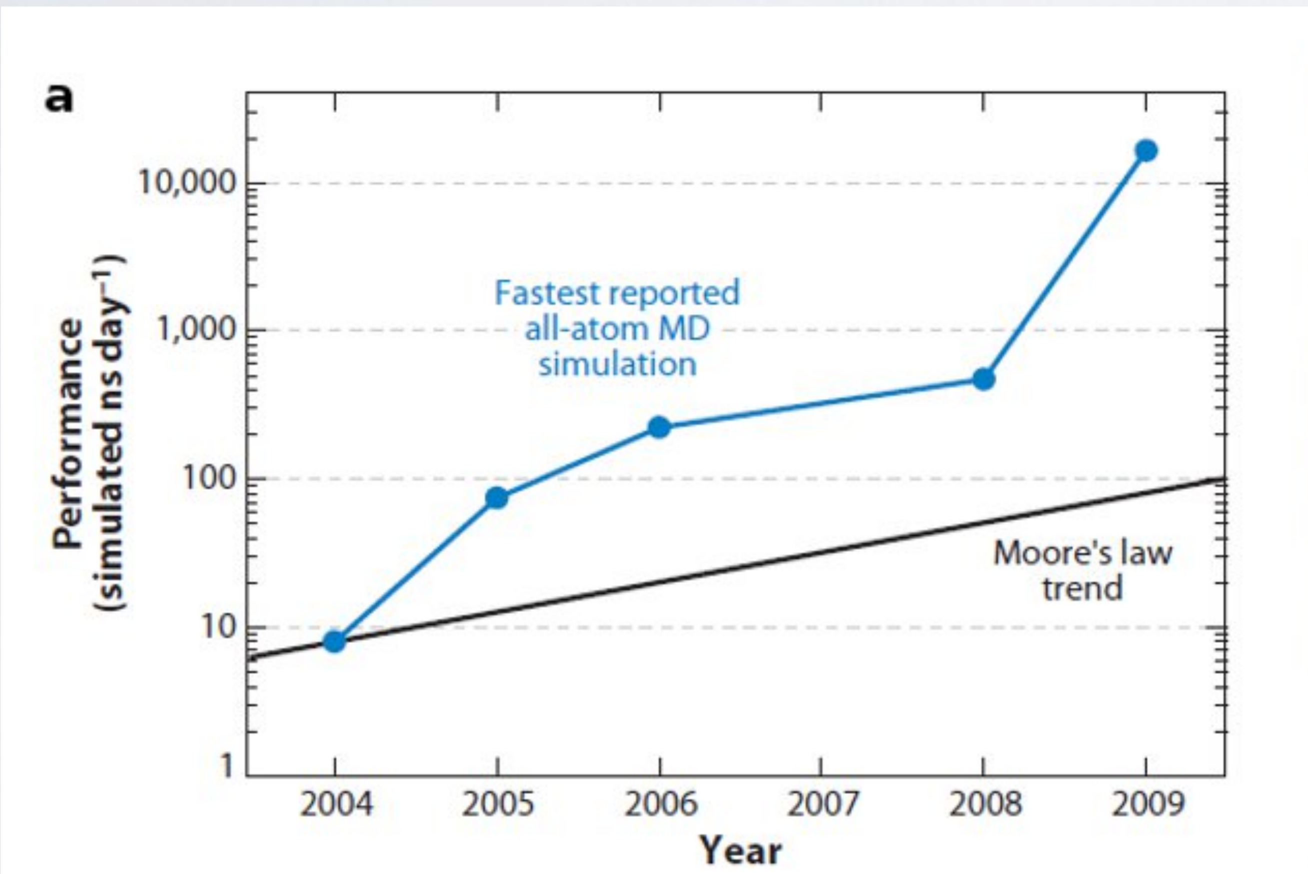
Useful, widely adopted but far from perfect
Multiple groups working on fewer, better approxs
Force fields, quantum
entropy, water effects
Moore's law: hardware improving


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



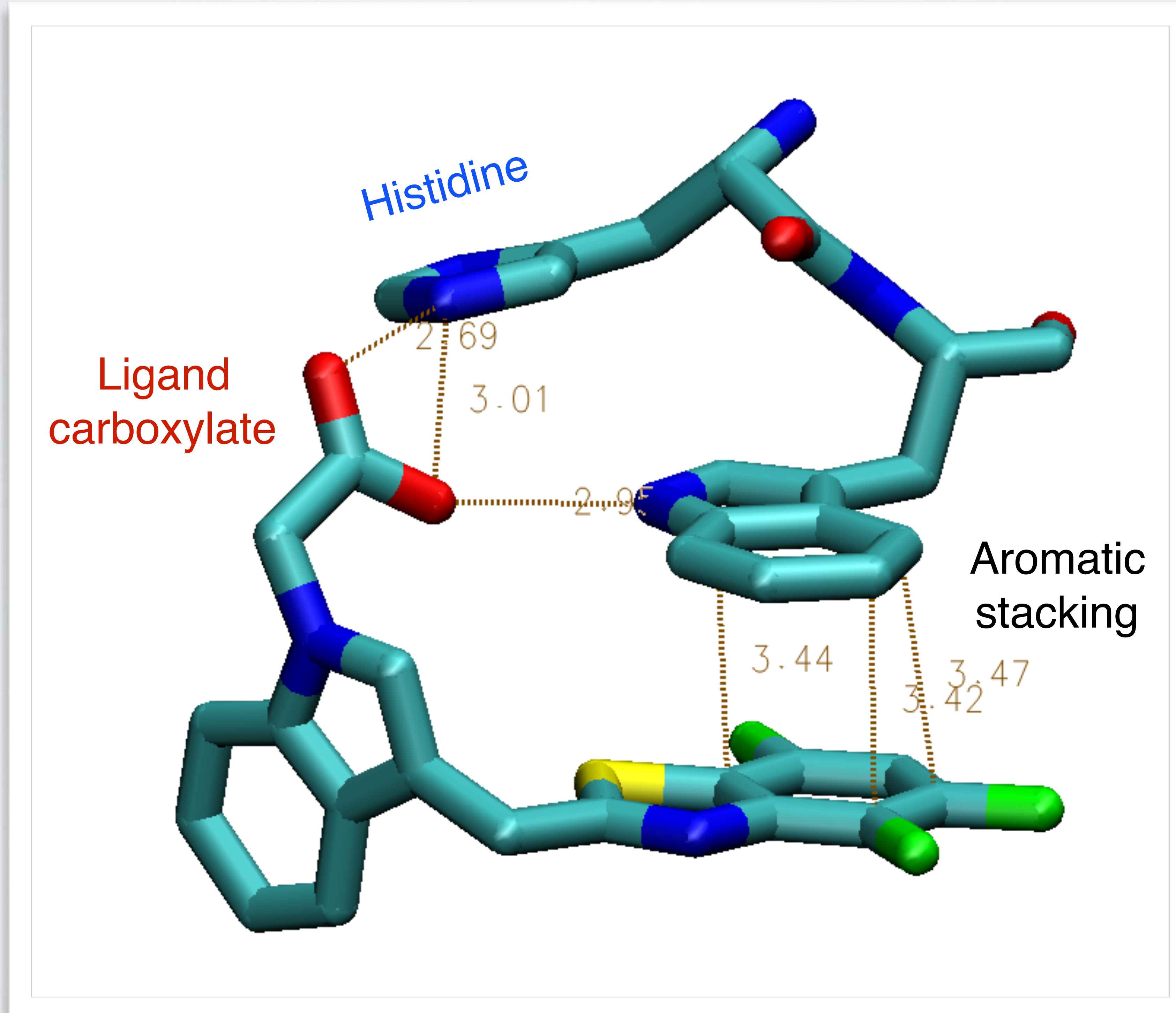
POTENTIAL FUNCTIONS DESCRIBE A SYSTEMS **ENERGY** AS A FUNCTION OF ITS **STRUCTURE**

Two main approaches:

(1). Physics-Based

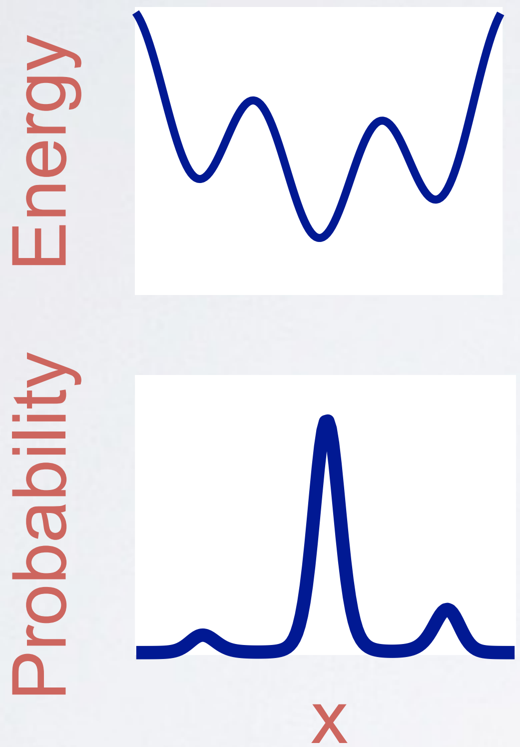
(2). Knowledge-Based

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:

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

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 POTENTIALS

Weaknesses

Accuracy limited by availability of data

Strengths

Relatively easy to implement
Computationally fast

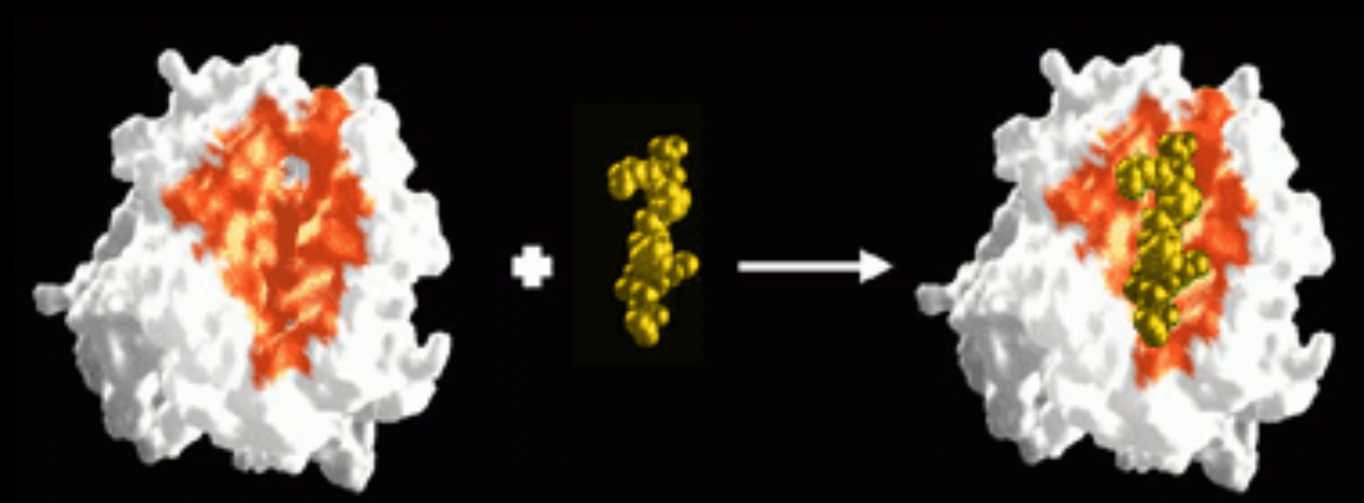
Status

Useful, far from perfect
May be at point of diminishing returns
(not always clear how to make improvements)

- Break -

Download MGL Tools: See class website!

Computer Aided 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 DRUG DISCOVERY

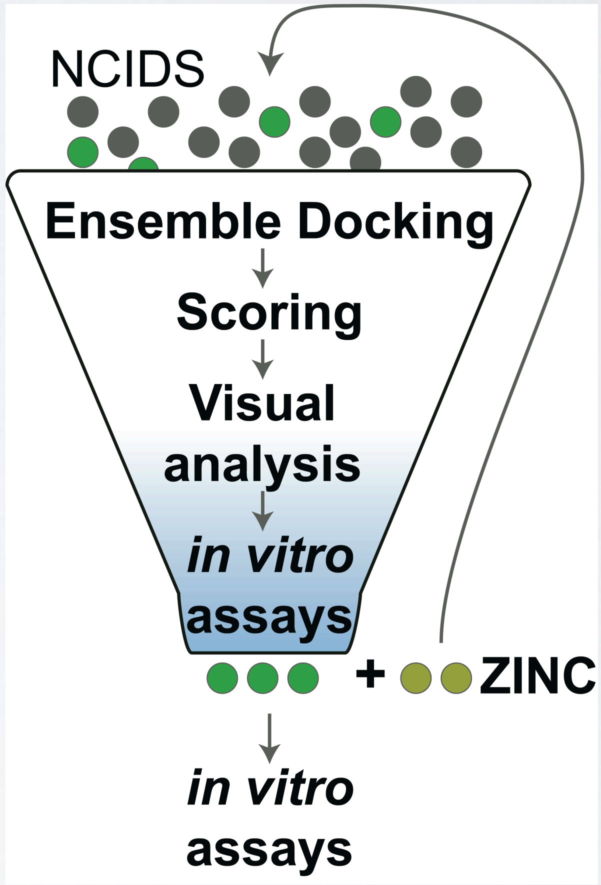
Aims to reduce number of compounds synthesized and assayed

Lower costs

Reduce chemical waste

Facilitate faster progress

N.B. Comparable experimental screens often out of reach of academia (facilities, cost)

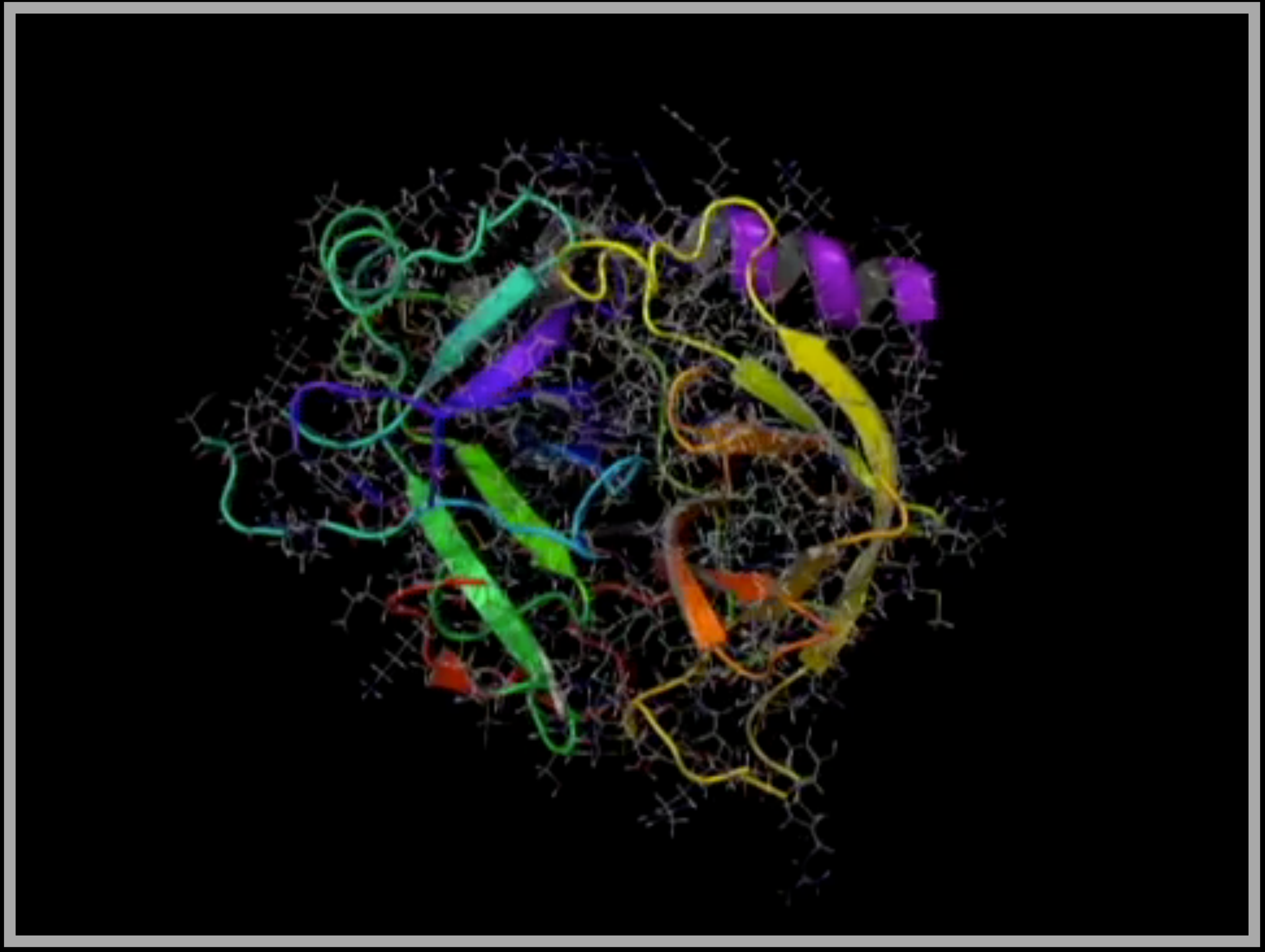


Applications...

- Discriminate between good and poor binders, or provide a priority ranking to a collection of ligands
- Provide in-depth mechanistic characterization of specific ligand or group of ligands
- Provide valuable guidance for medicinal chemists trying to synthesize ligands with improved properties (affinities and potencies)

Q. “How can we modify an already active ligand to make it even more active?”

Computational Ligand Docking



- Screening and ranking compounds as potential ligands (a.k.a. virtual screening)
- Improving "lead" compounds (a.k.a. ligand optimization, more on this later...)
  - This is a common practice among seasoned computational chemists

Two main approaches:

(1). Receptor/Target-Based

(2). Ligand/Drug-Based

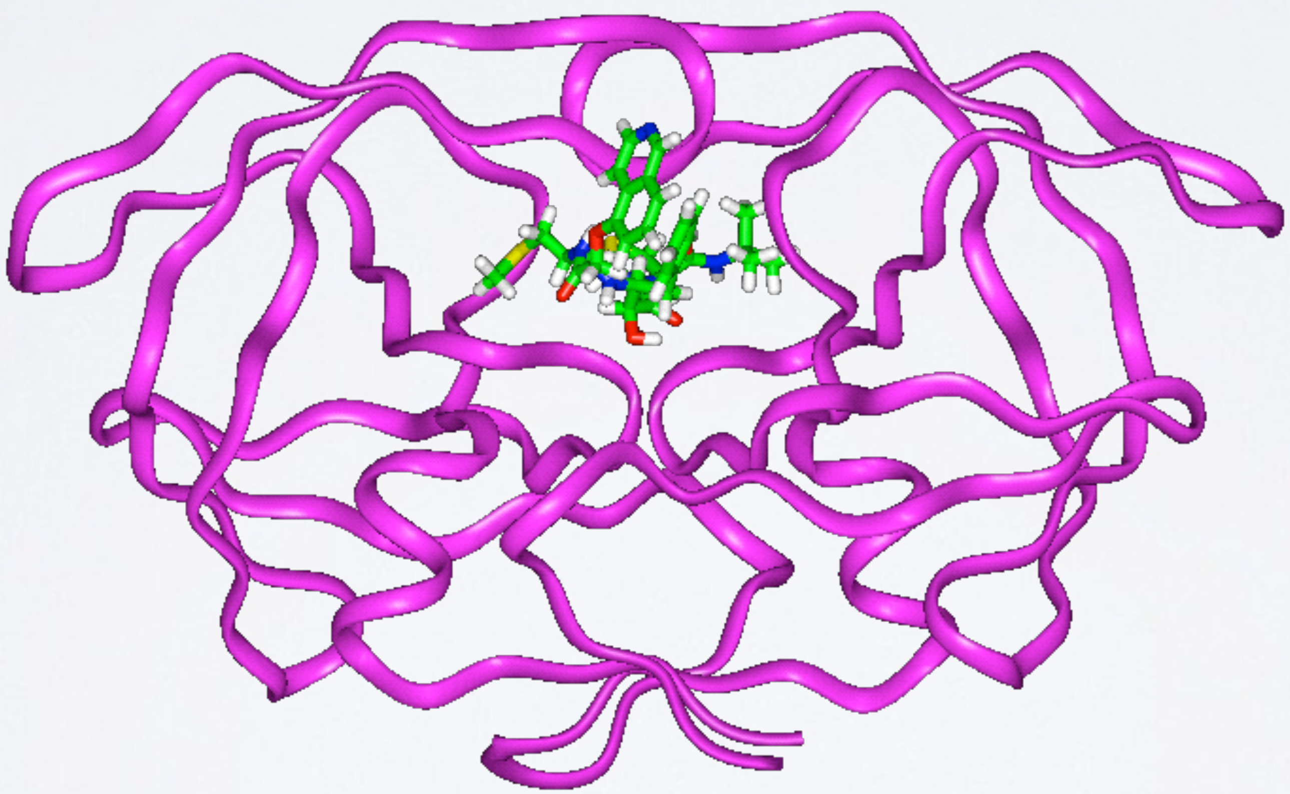
Two main approaches:

(1). Receptor/Target-Based

(2). Ligand/Drug-Based

SCENARIO 1:
RECEPTOR-BASED DRUG DISCOVERY

Structure of Targeted Protein Known: Structure-Based Drug Discovery



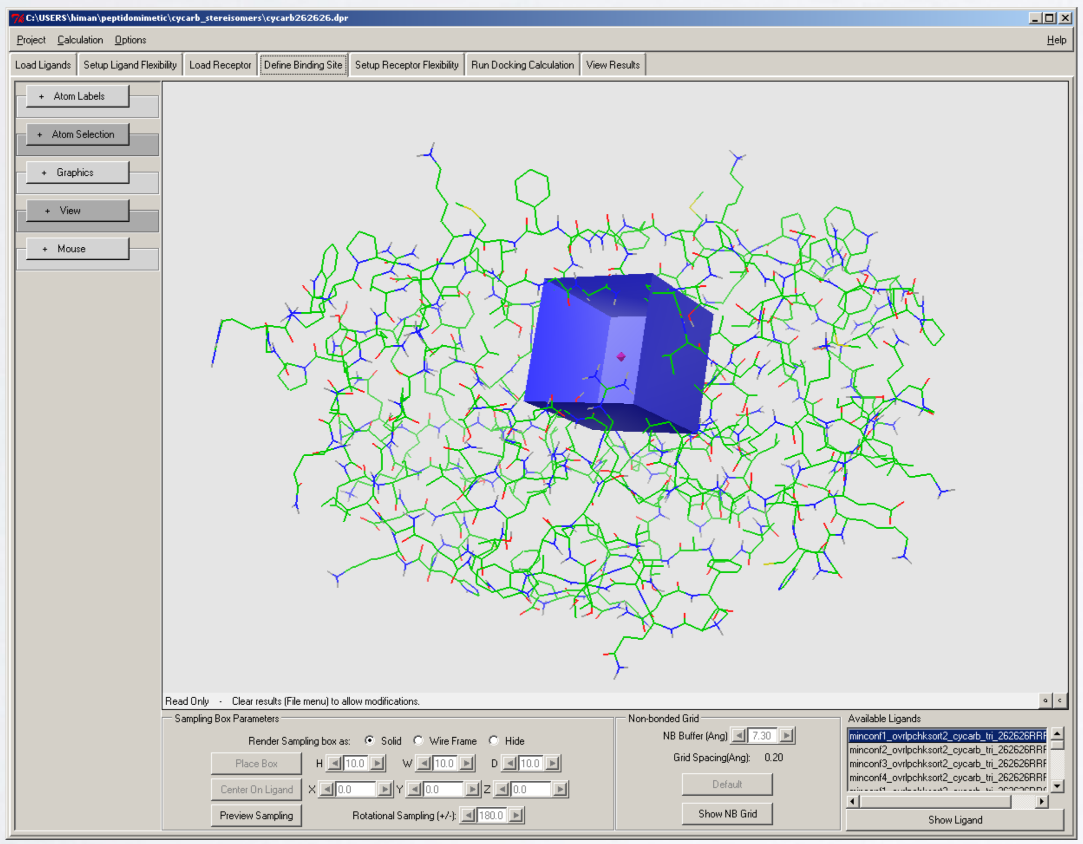
HIV Protease/KNI-272 complex

PROTEIN-LIGAND DOCKING

Structure-Based Ligand Design

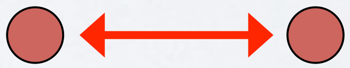
Docking software

Search for structure of lowest energy

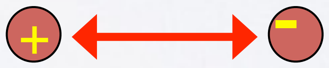


Potential function

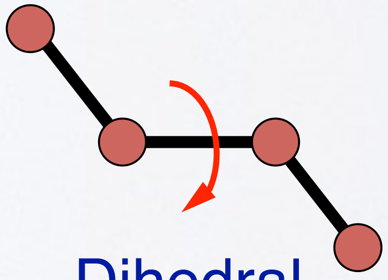
Energy as function of structure



VDW

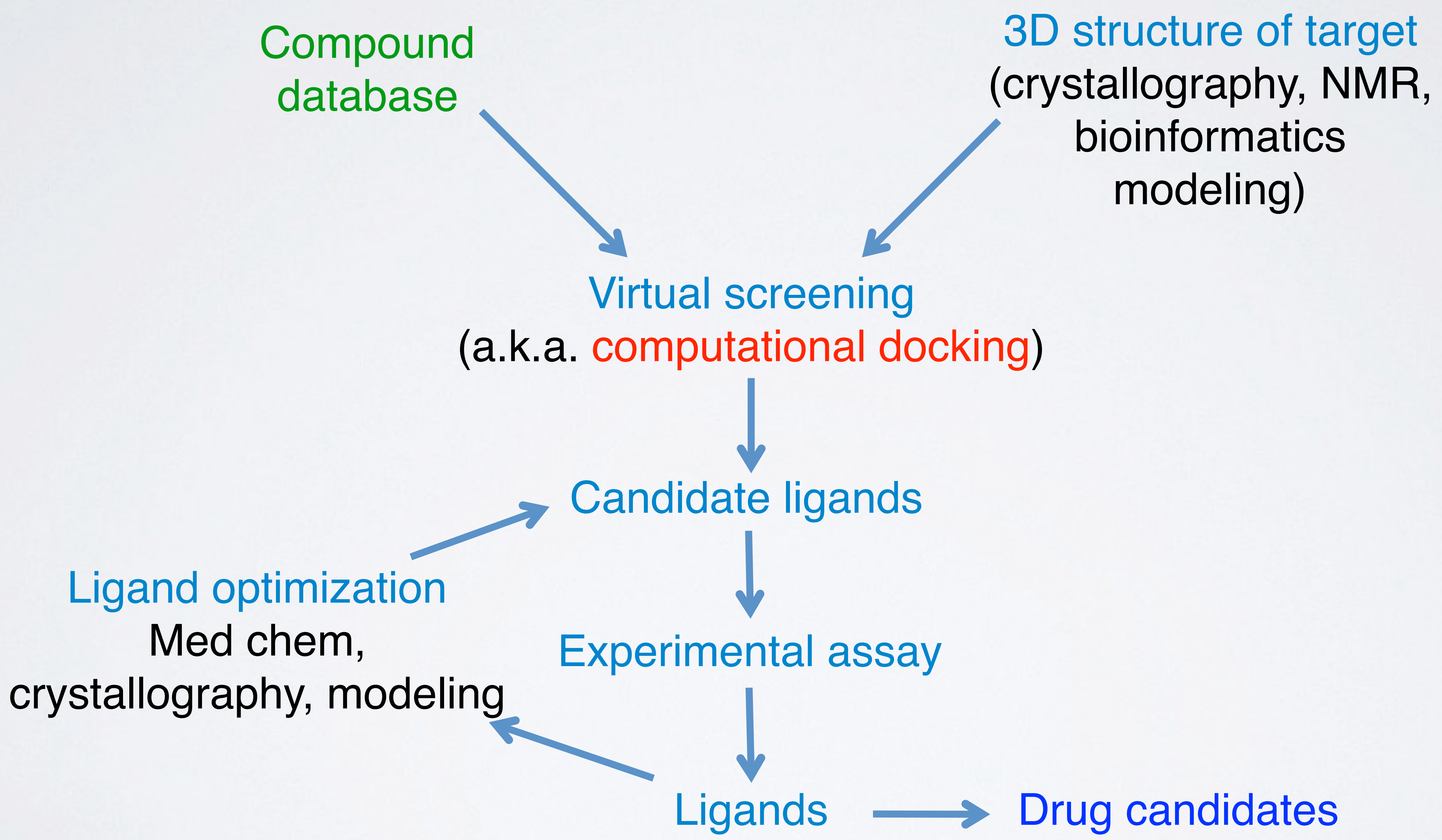


Screened Coulombic

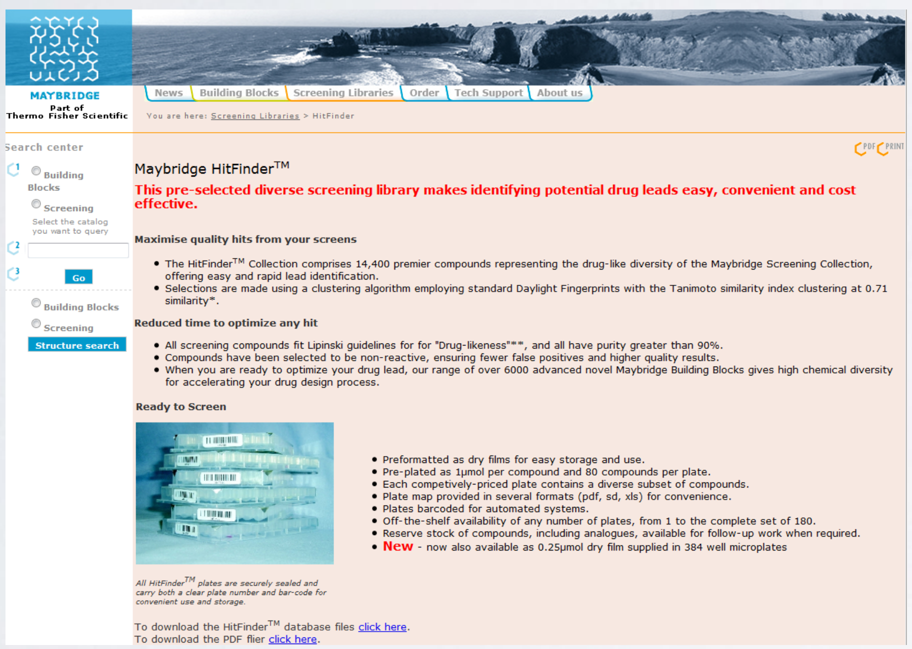


Dihedral

STRUCTURE-BASED VIRTUAL SCREENING



COMPOUND LIBRARIES

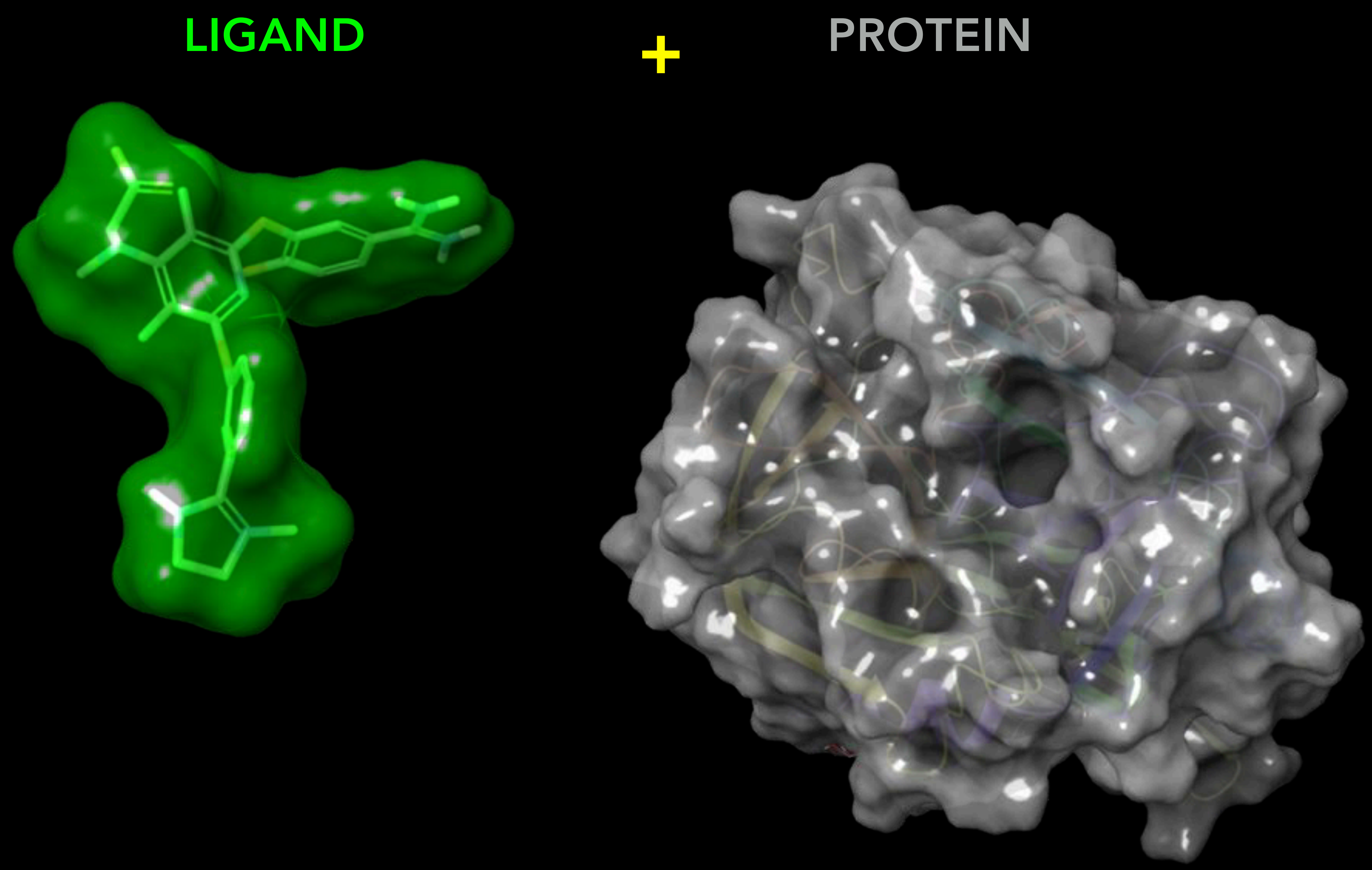


Commercial
(in-house pharma)

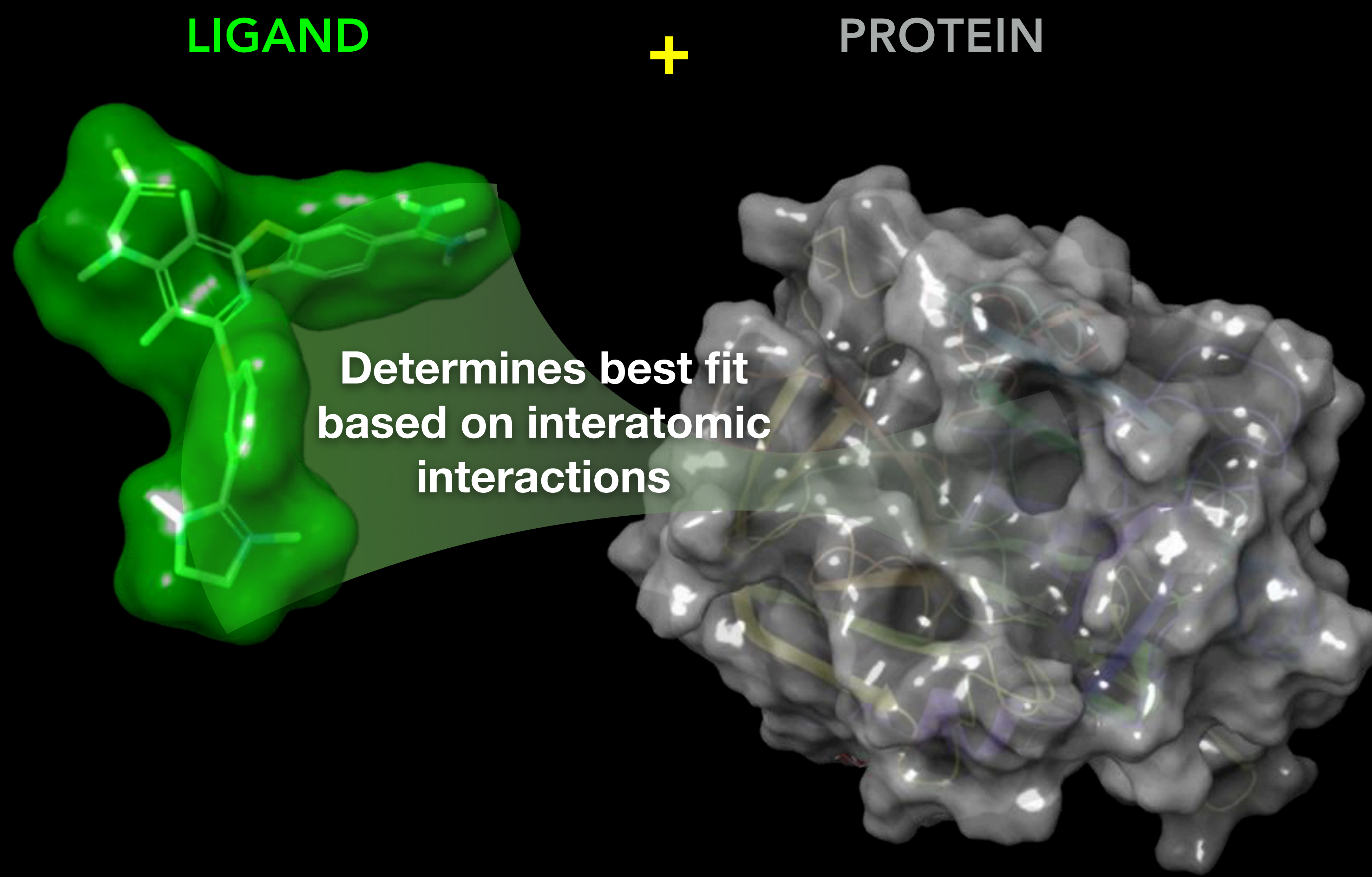
Government (NIH)

Academia

Docking at its core is a shape matching problem



Docking at its core is a shape matching problem



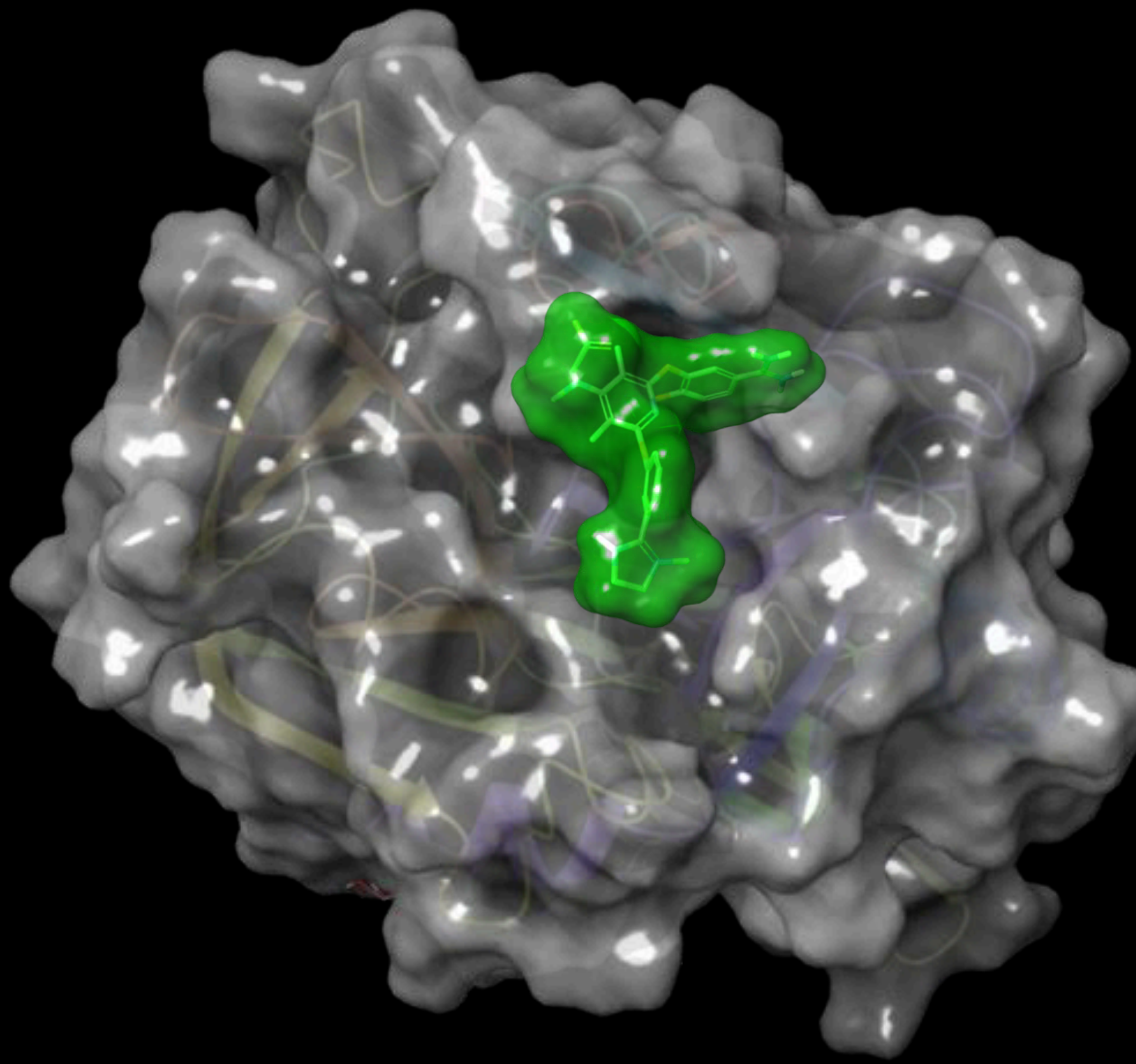
$$V(R) = E_{bonded} + E_{non.bonded}$$

Bonding Interactions

- Bond length
- Bond angels
- Torsions

Non-Bonding Interactions

- van der Waal's interactions
- H-bonds
- Charge-Charge interactions
- pi-pi, pi-cation, etc.



PROTEIN-LIGAND
complex

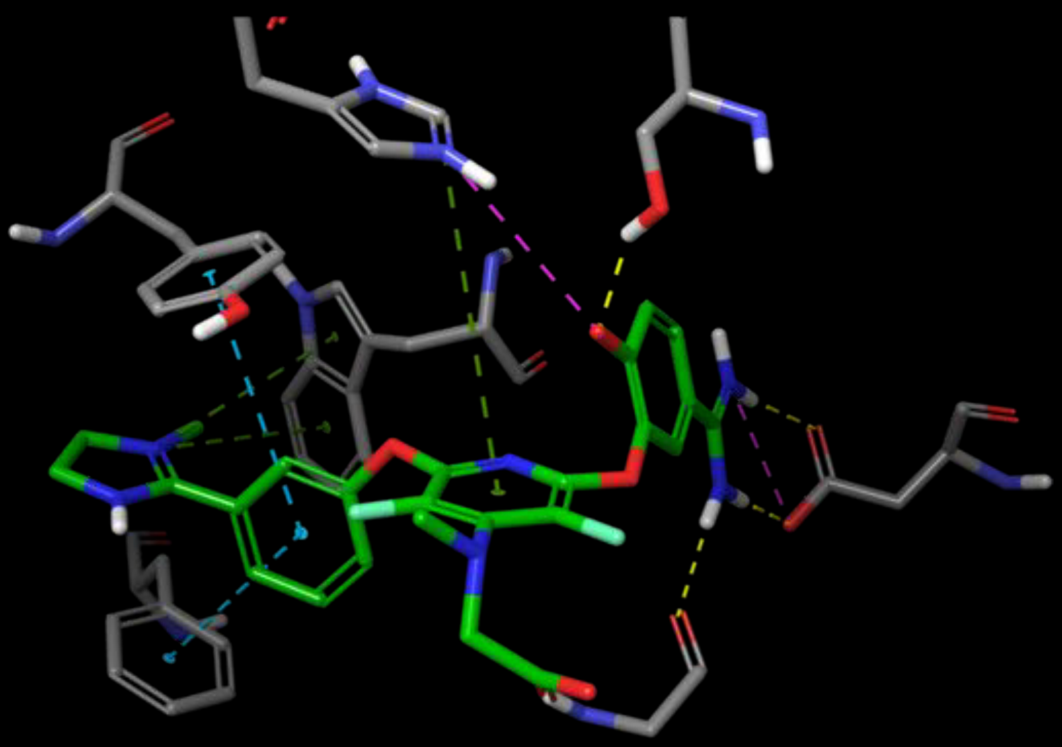
Do it Yourself!

Hand-on time!

http://thegrantlab.org/bggn213/

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

A Docking Program Generates a...

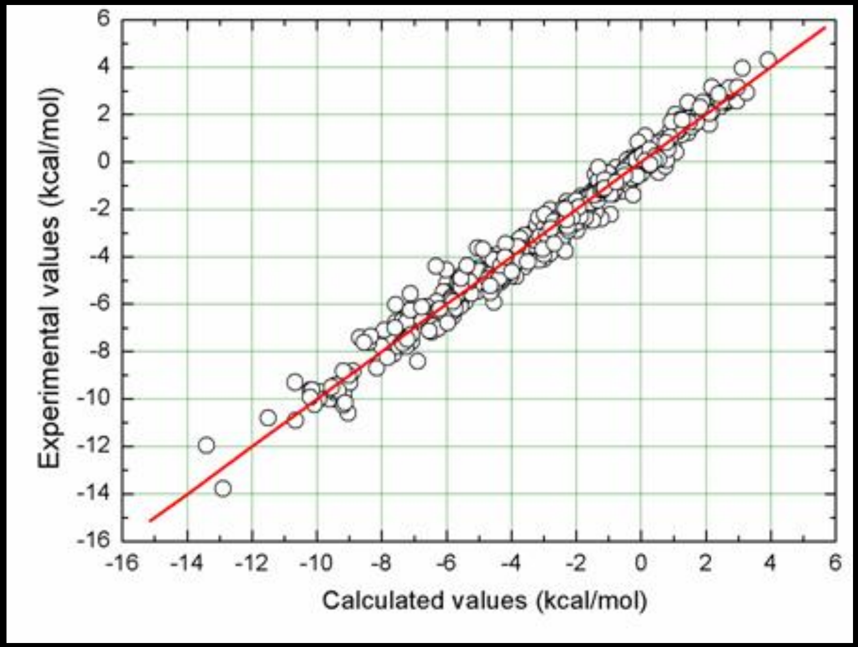


1. **Binding Pose**
   A model of the ordination of the ligand in the binding site of the receptor.

2. **Docking Score**
   A numerical value representing the quality of the pose. Often presented as binding energy.

Scoring functions enable different docking results to be compared

- **Scoring functions aim to estimate ligand binding affinity, or the free energy of binding (ΔG), so that different poses can be compared**
  - The posses with the most negative values are predicted to have the tightest interactions
- **Scoring functions are constructed from a weighted sum of all possible molecular interactions that contribute to binding**
  - Including H-bonds, van der Waals forces, electrostatic interactions, etc. and penalties for steric clashes and loss of entropy
- **Scoring systems are optimized and validated by fitting to experimental values for known receptor-ligand interactions**



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

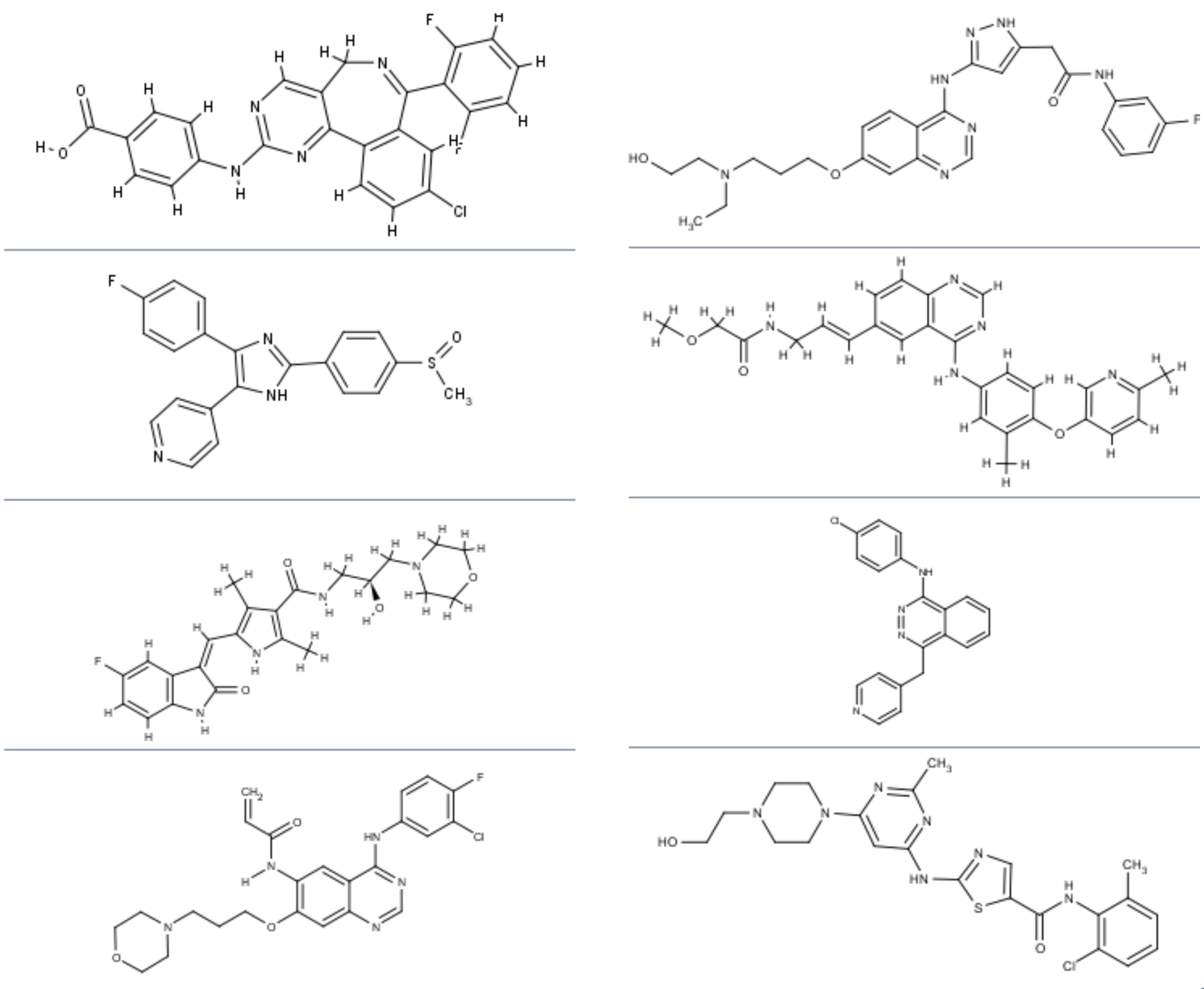
(2). Ligand/Drug-Based

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?

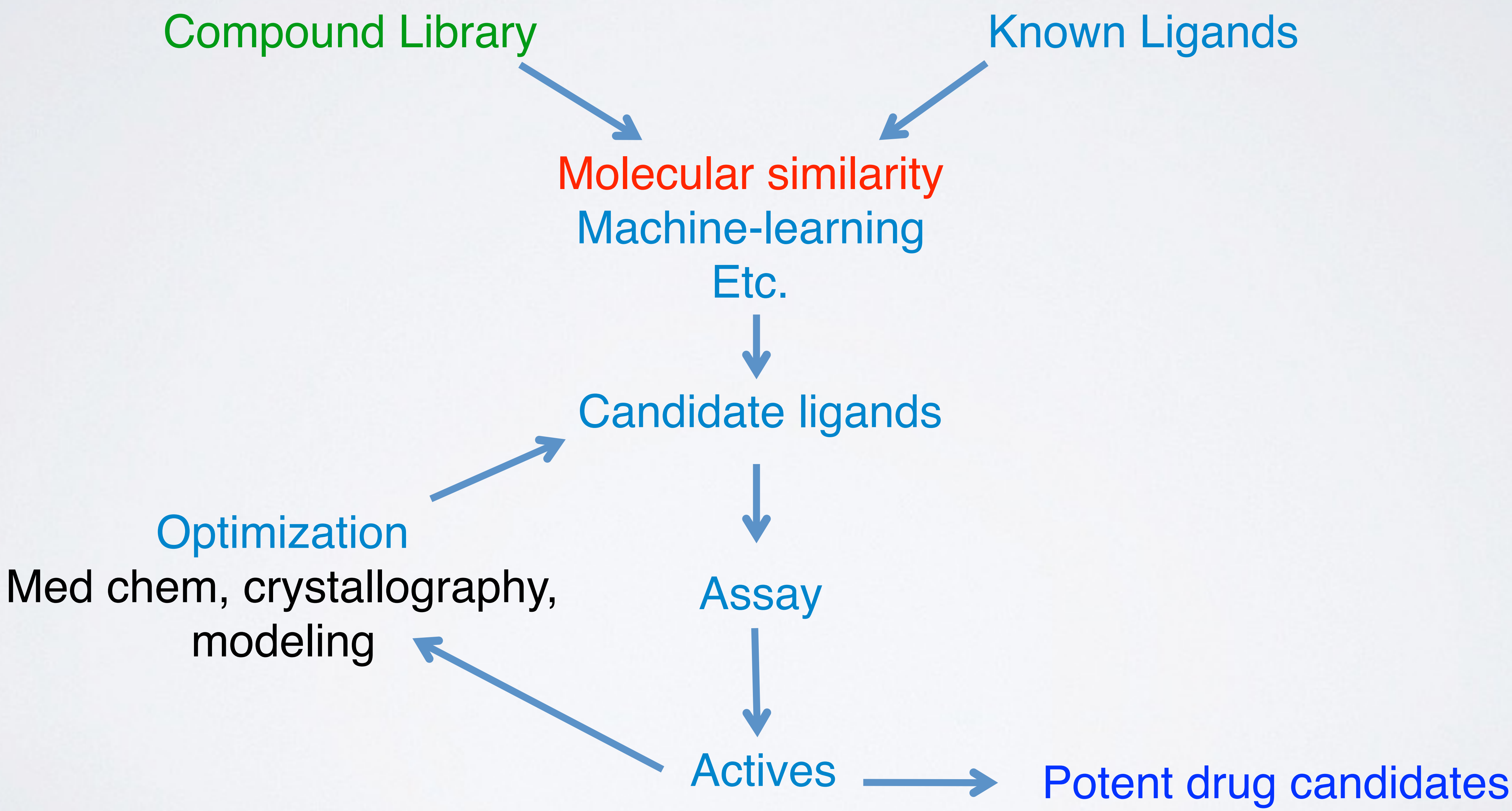
Experimental screening generated some ligands, but they don't bind tightly enough

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

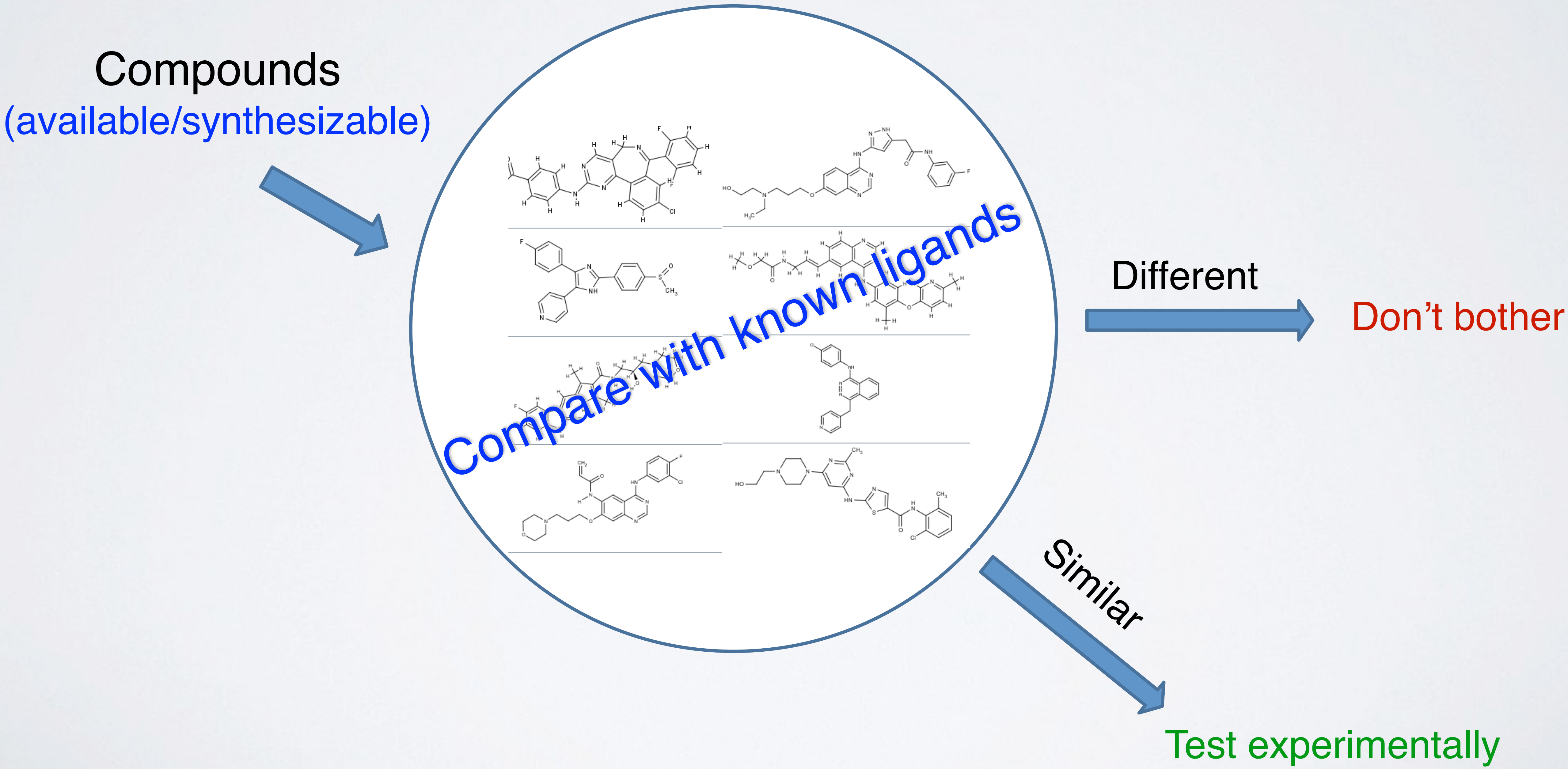
An high-affinity ligand is toxic, is not well-absorbed, difficult to synthesize etc.

Drug resistance variants of the receptor have emerged...

LIGAND-BASED VIRTUAL SCREENING

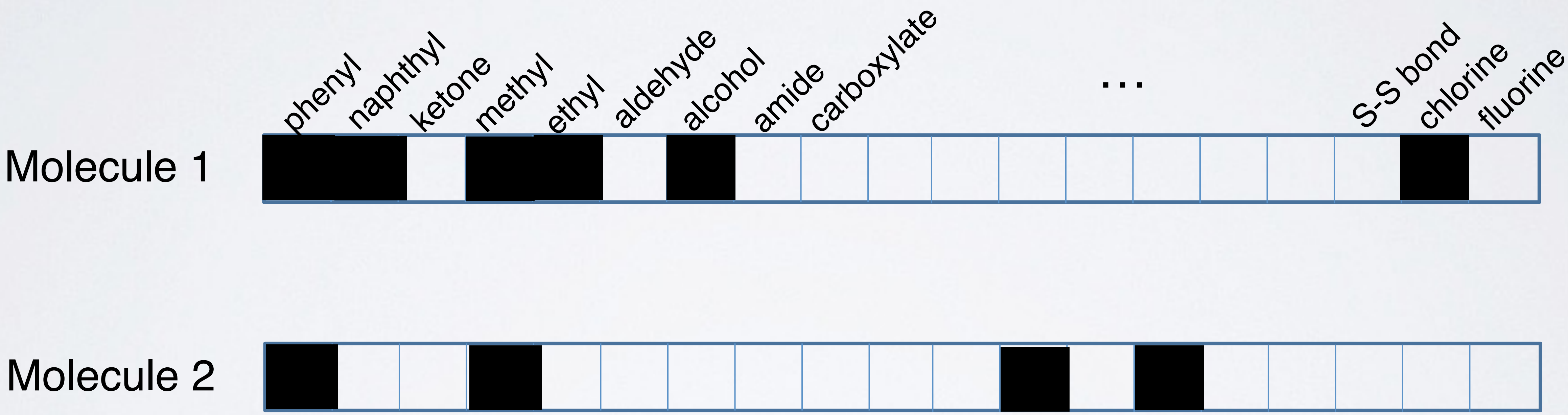


CHEMICAL SIMILARITY
LIGAND-BASED DRUG-DISCOVERY

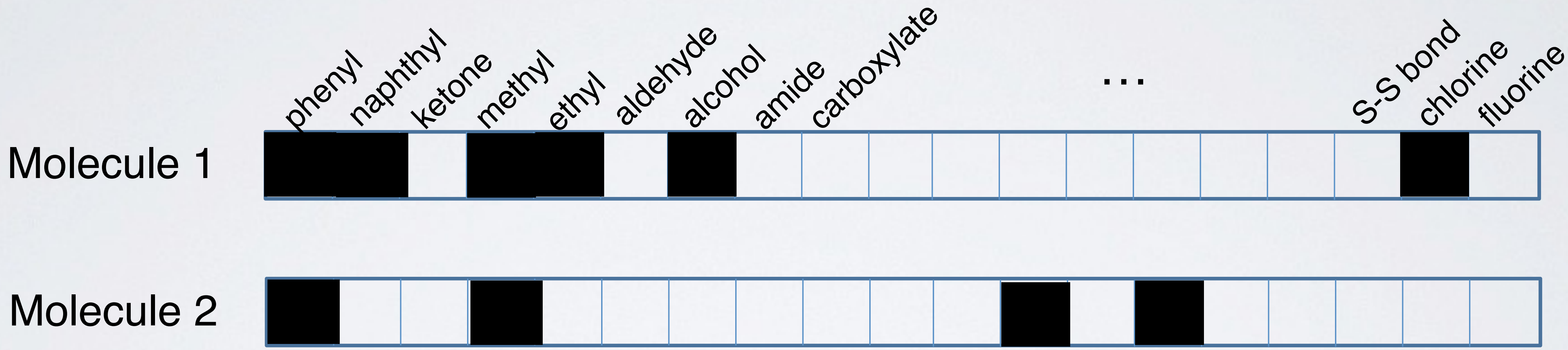


CHEMICAL FINGERPRINTS

BINARY STRUCTURE KEYS



CHEMICAL SIMILARITY FROM FINGERPRINTS



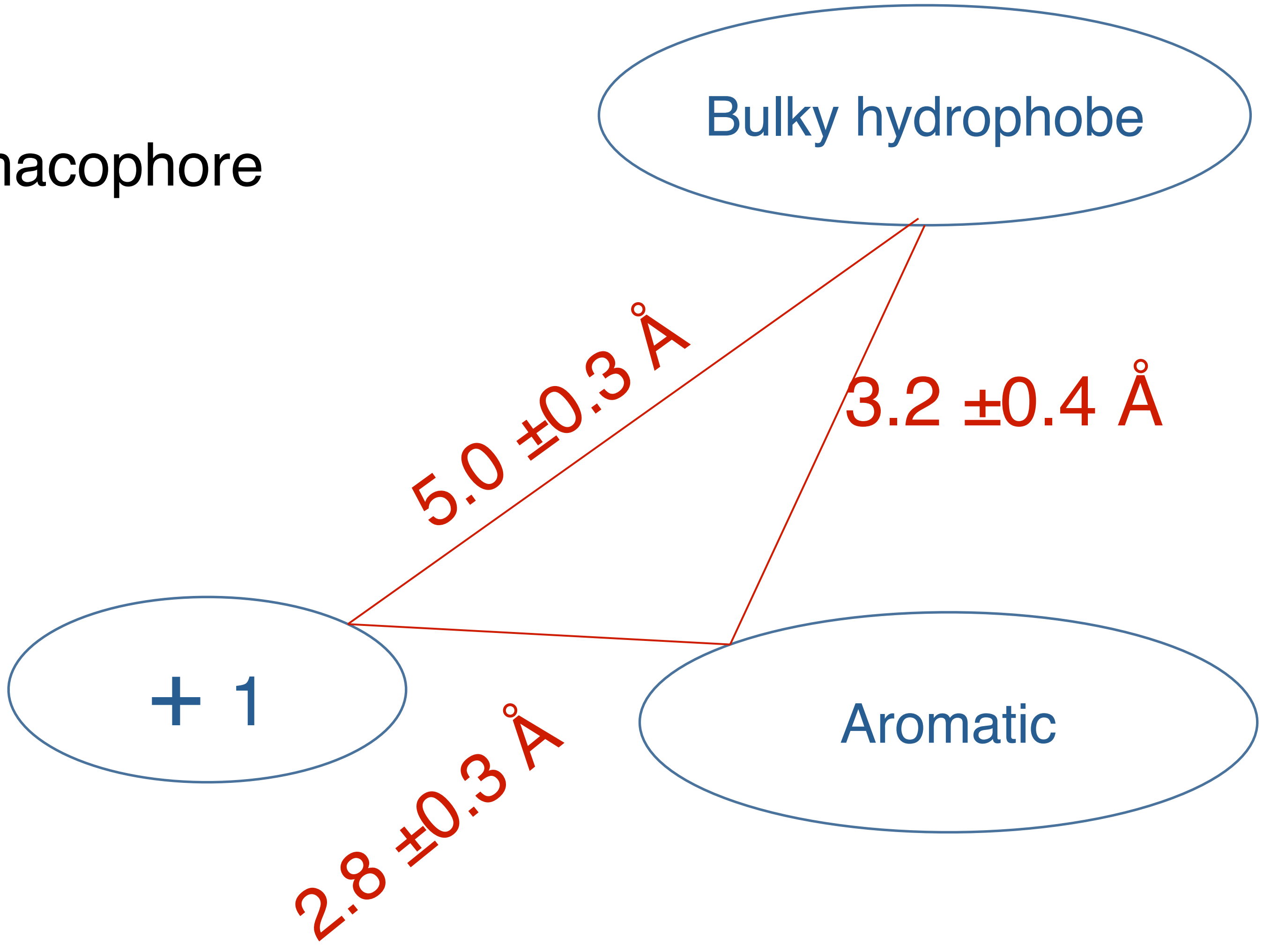
Tanimoto Similarity (or Jaccard Index), T

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

Pharmacophore Models

Φάρμακο (drug) + Φορά (carry)

A 3-point pharmacophore

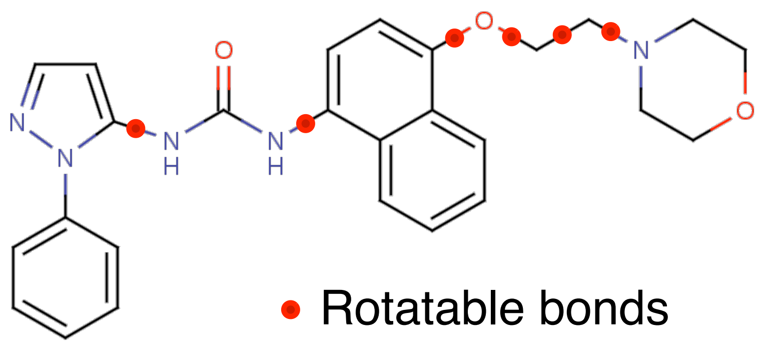


Molecular Descriptors

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

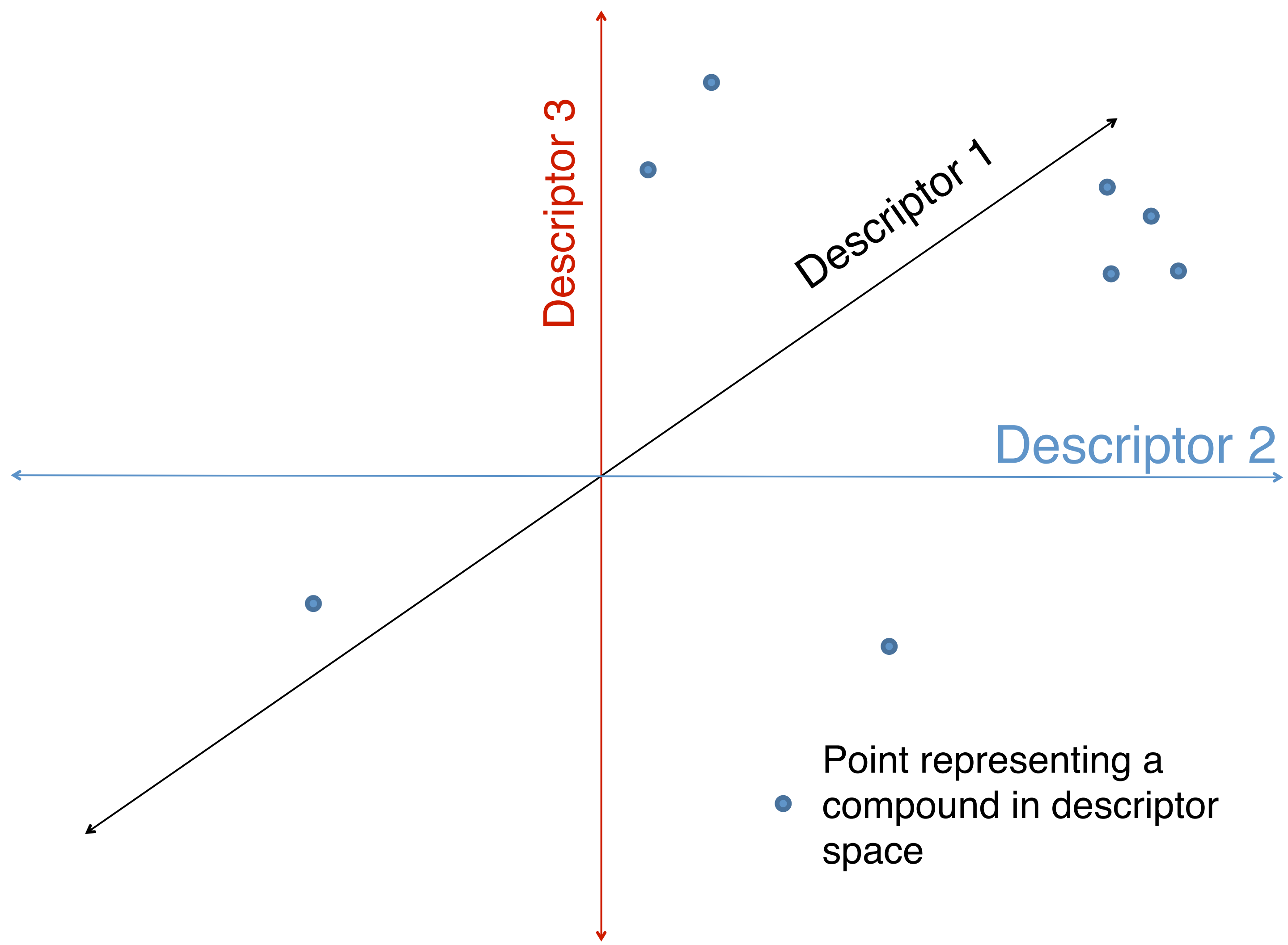
- branching index
- measures of linearity vs interconnectedness

Etc. etc.

A High-Dimensional "Chemical Space"

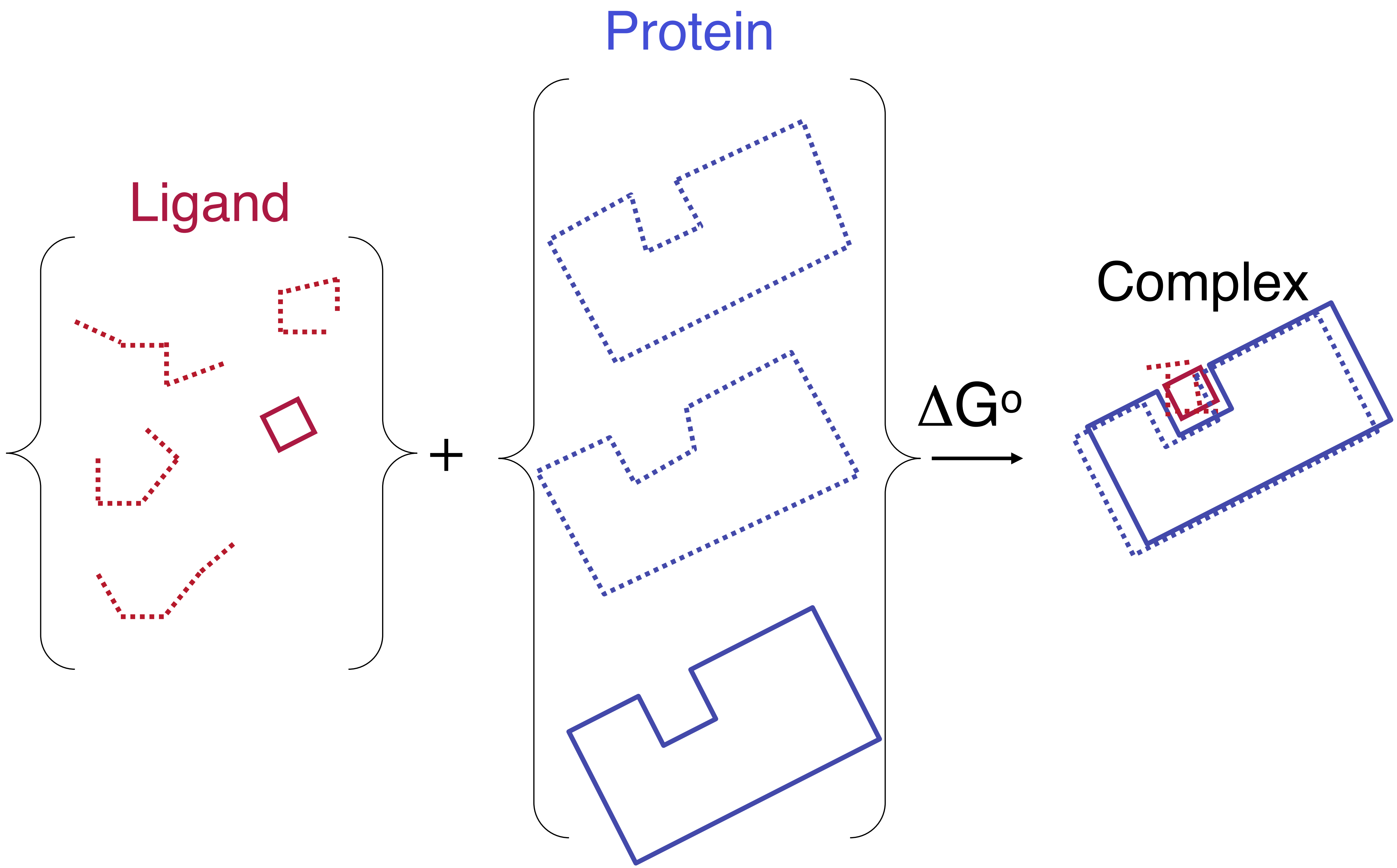
Each compound is a point in an n-dimensional space

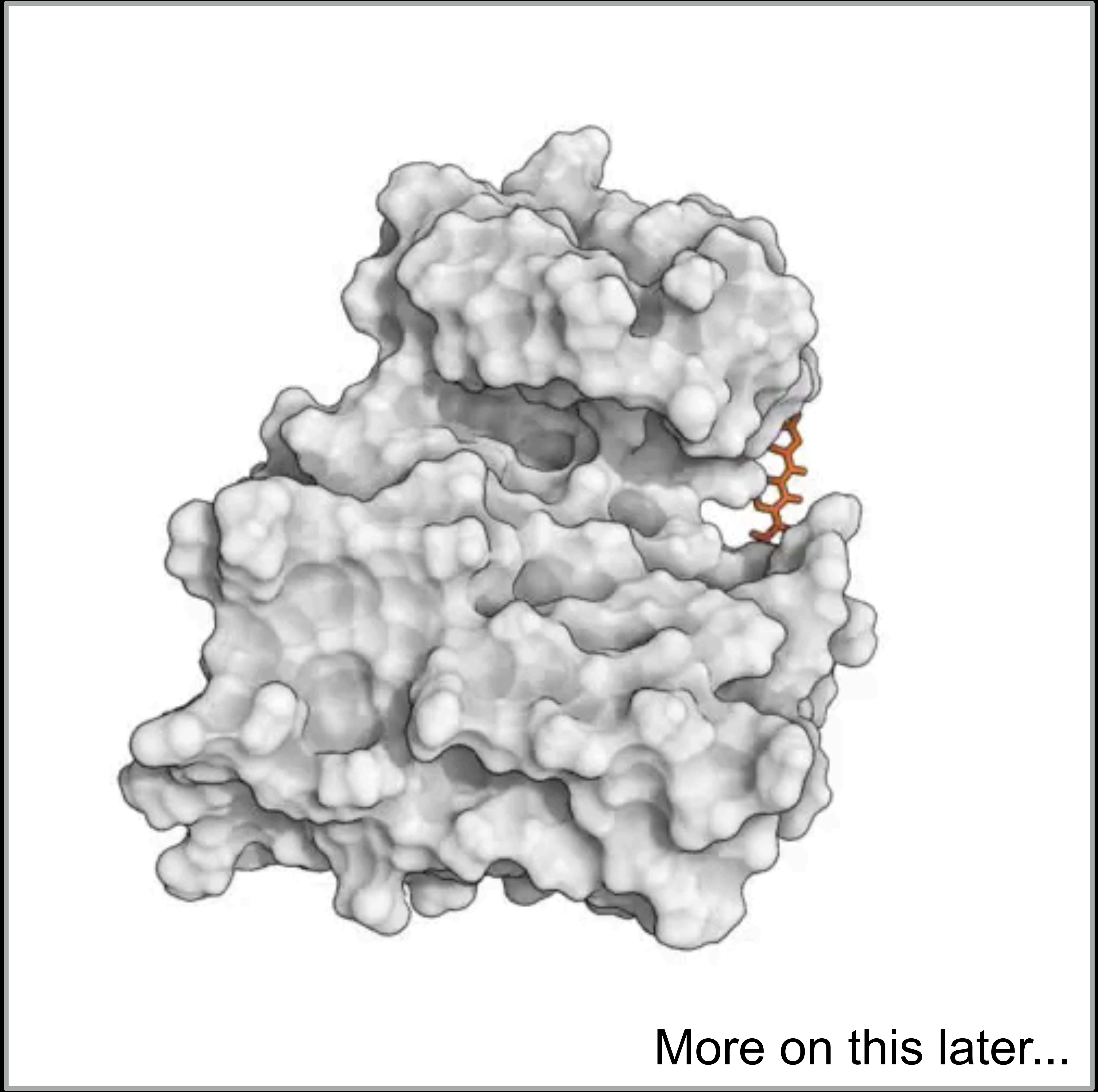
Compounds with similar properties are near each other



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

Key Challenge: Proteins & Ligand are Flexible

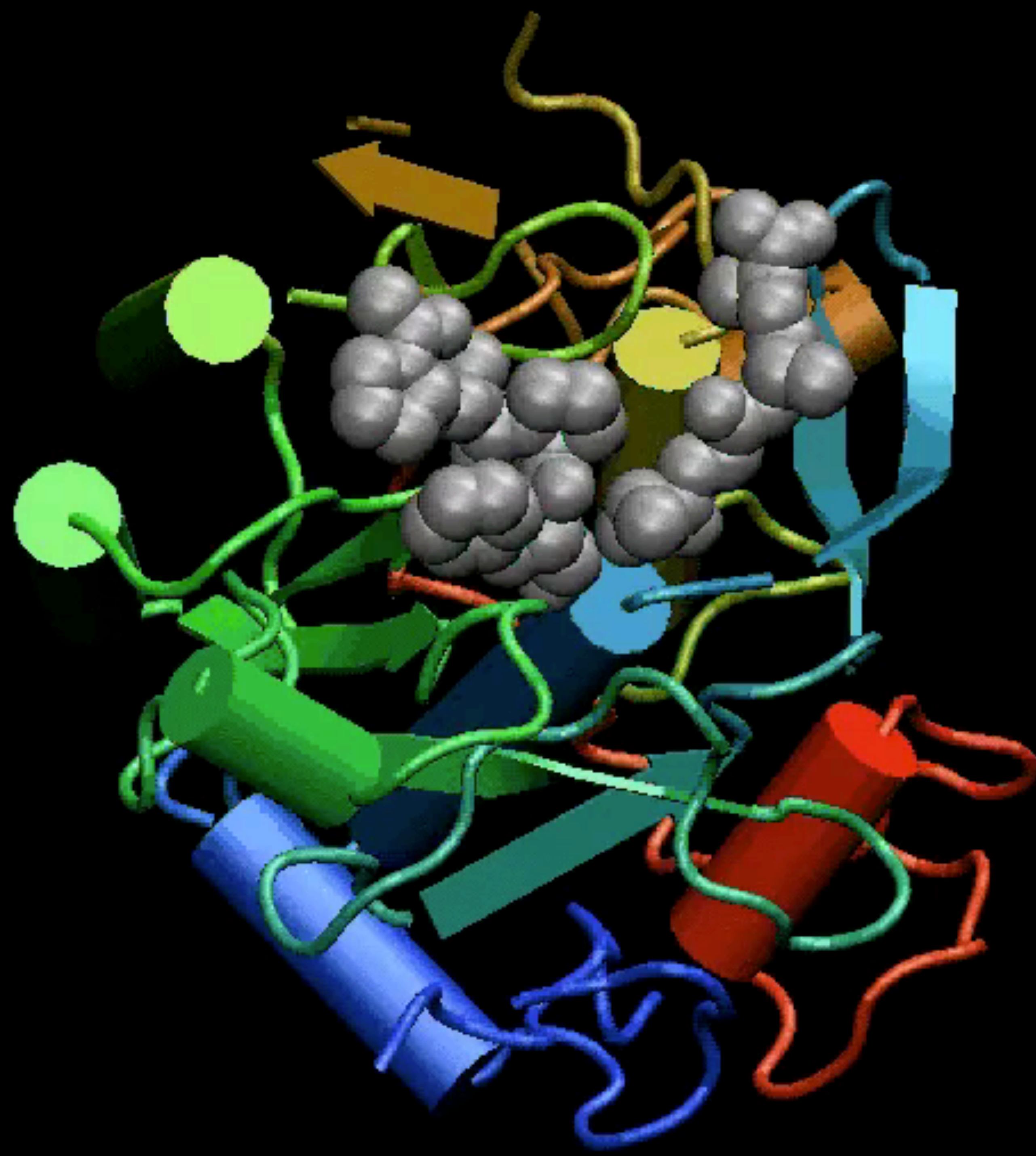




Proteins are flexible, which is a limitation in current rigid docking approaches… but when combined with molecular dynamics bioinformatics can be a powerful tool!

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

Do it Yourself!



https://bioboot.github.io/bggn213_F19/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)))
```

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

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.

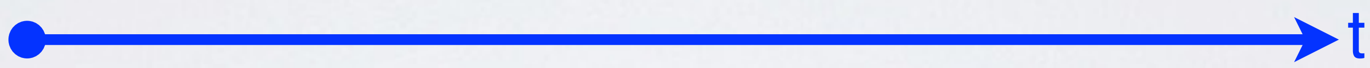
- Thus knowledge of dynamics can provide a deeper understanding of the mapping of structure to function
  - Molecular dynamics (MD) and normal mode analysis (NMA) are two major methods for predicting and characterizing molecular motions and their properties

MOLECULAR DYNAMICS SIMULATION

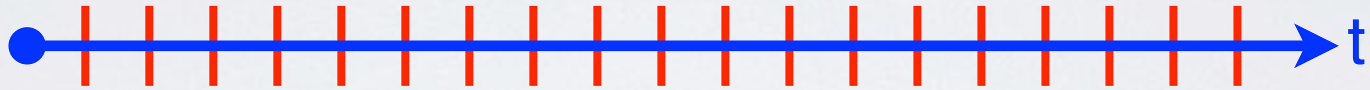
- Use force-field to find Potential energy between all atom pairs
- Move atoms to next state
- Repeat to generate trajectory

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

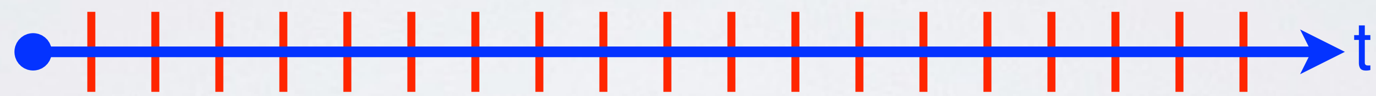
- Divide time into discrete (~1fs) time steps (Δt)
(for integrating equations of motion, see below)



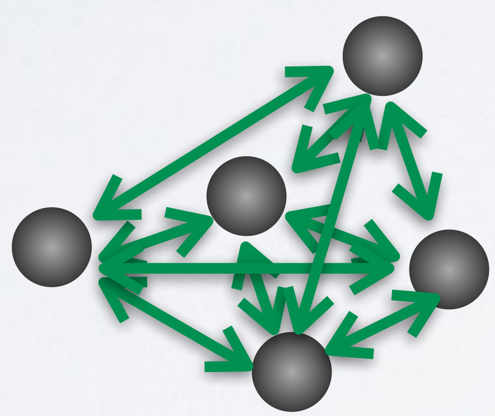
- Divide time into discrete (~1fs) time steps (Δt)
(for integrating equations of motion, see below)



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



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



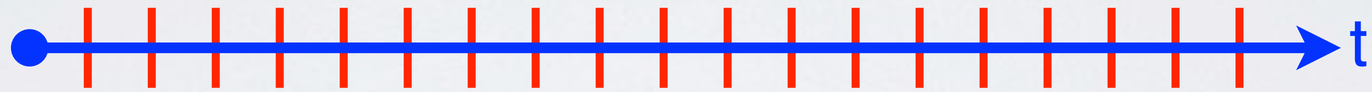
Nucleic motion described classically

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

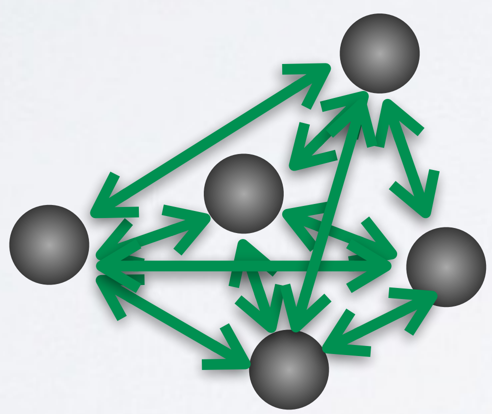
Empirical force field

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

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



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



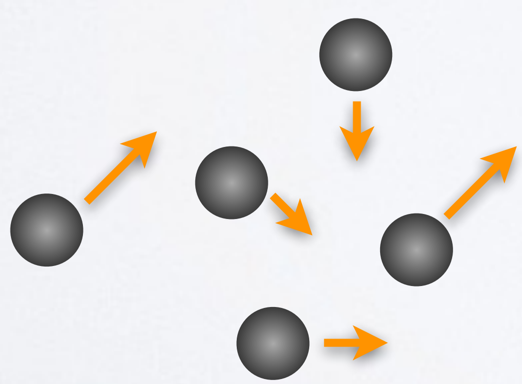
Nucleic motion described classically

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

Empirical force field

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

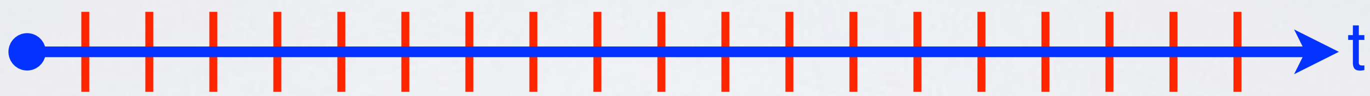
- Use the forces to calculate velocities and move atoms to new positions
(by integrating numerically via the "leapfrog" scheme)



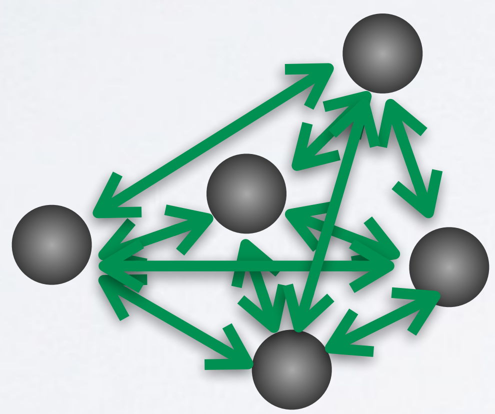
$$\begin{aligned} \boxed{\boldsymbol{v}(t+\frac{\Delta t}{2})} &= \boldsymbol{v}(t-\frac{\Delta t}{2}) + \frac{\boldsymbol{F}(t)}{m}\Delta t \\ \boldsymbol{r}(t+\Delta t) &= \boldsymbol{r}(t) + \boxed{\boldsymbol{v}(t+\frac{\Delta t}{2})}\Delta t \end{aligned}$$

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

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

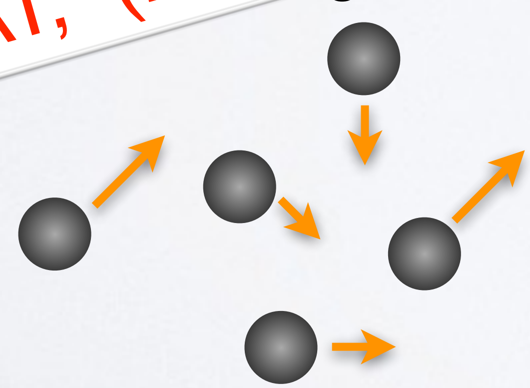
Empirical force f

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

non−bonded

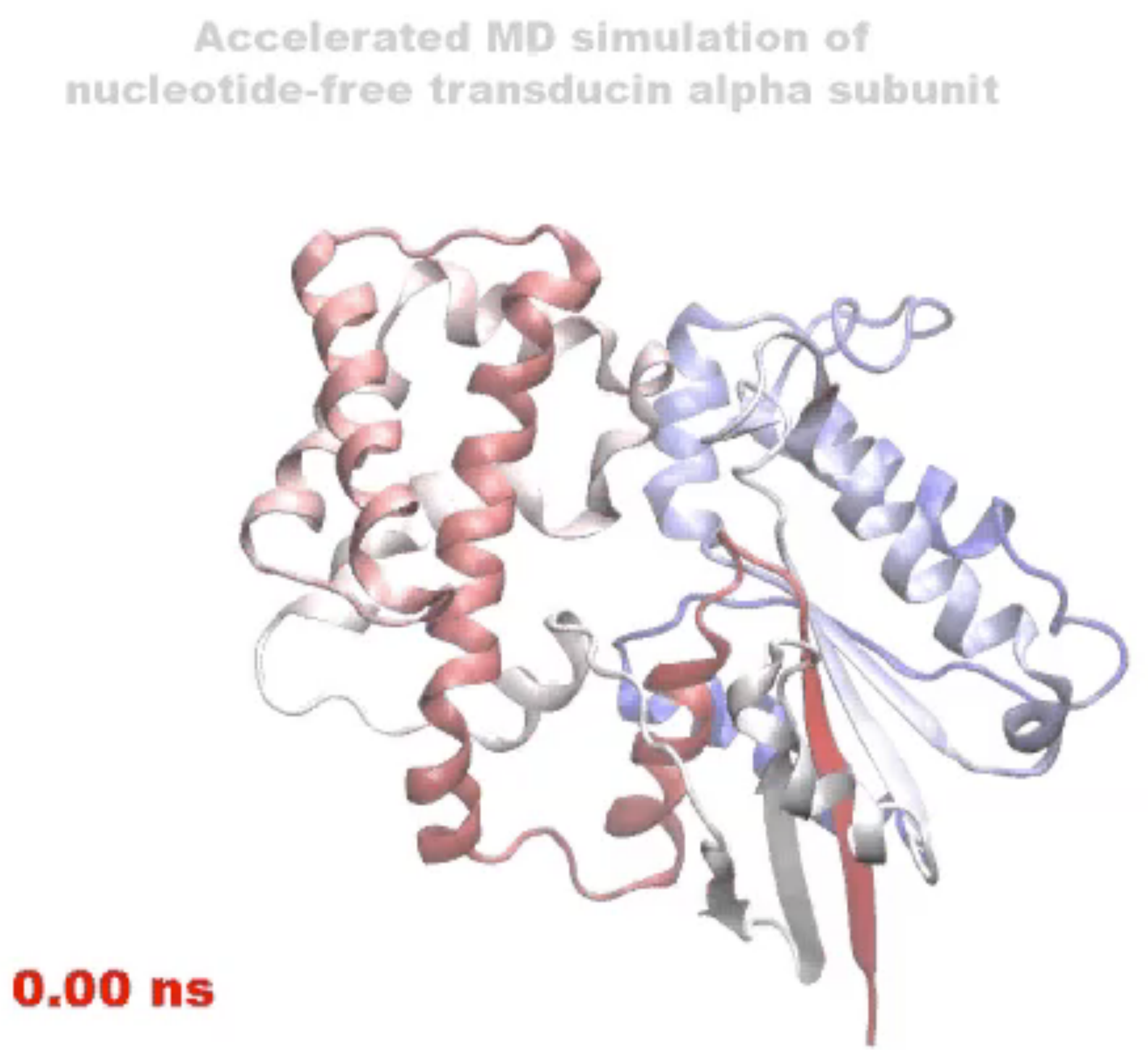
- REPEAT, (iterate many, many times… 1ms = 10^{12} time steps)

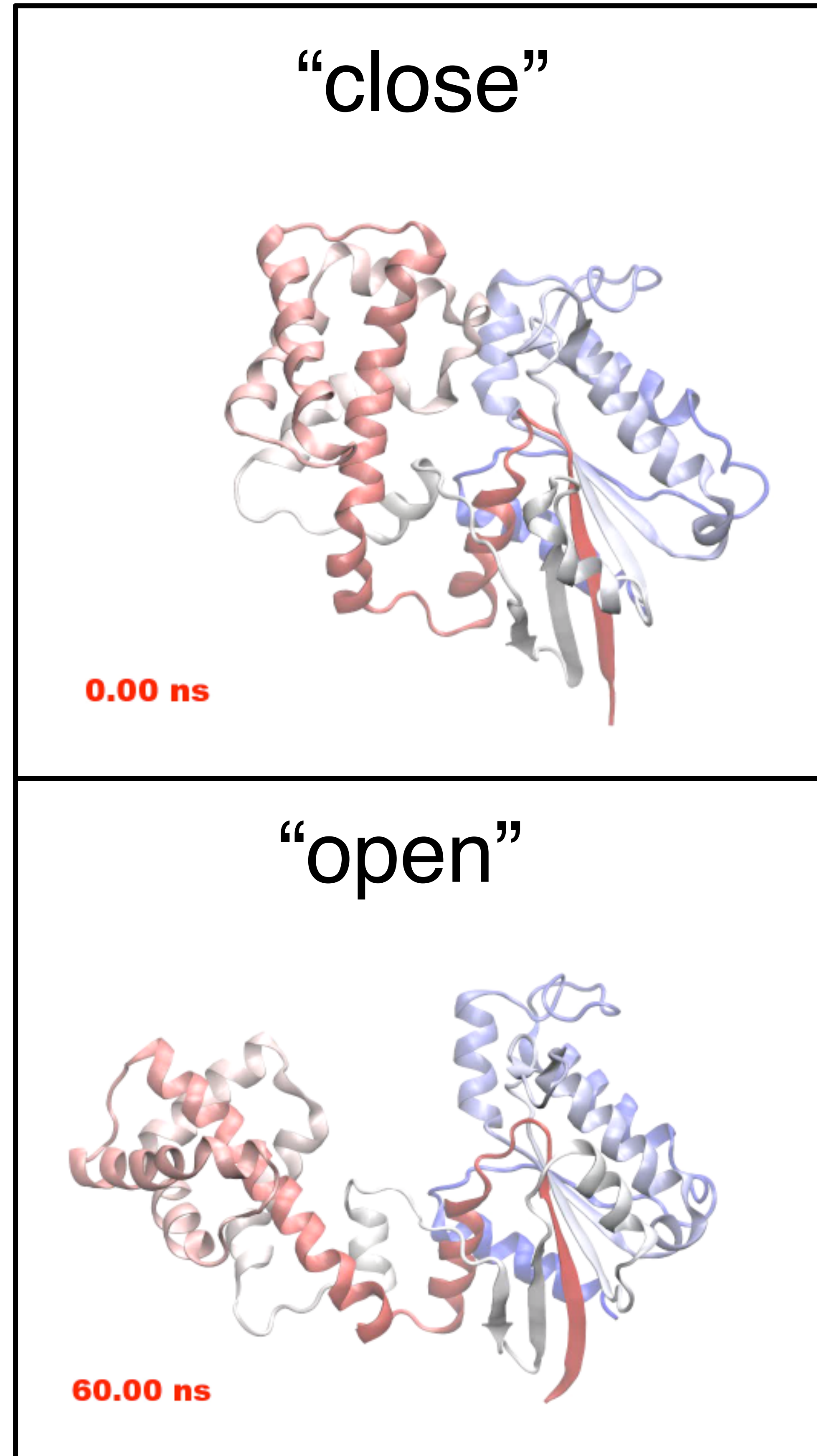
- Use the fo ate velocities and move atoms to new positions
  (b g numerically via the “leapfrog” scheme)



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

MD Prediction of Functional Motions

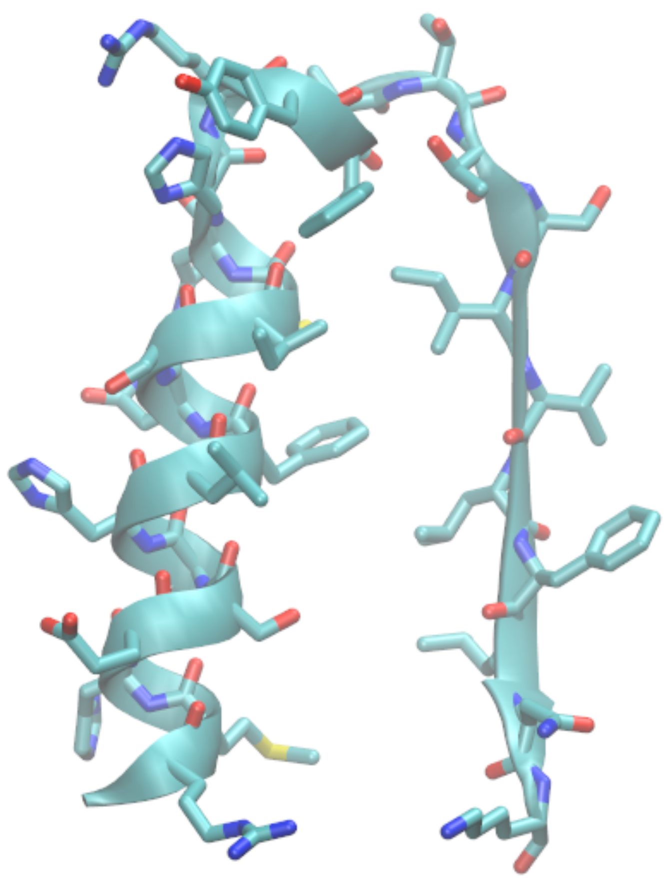




Yao and Grant, Biophys J. (2013)

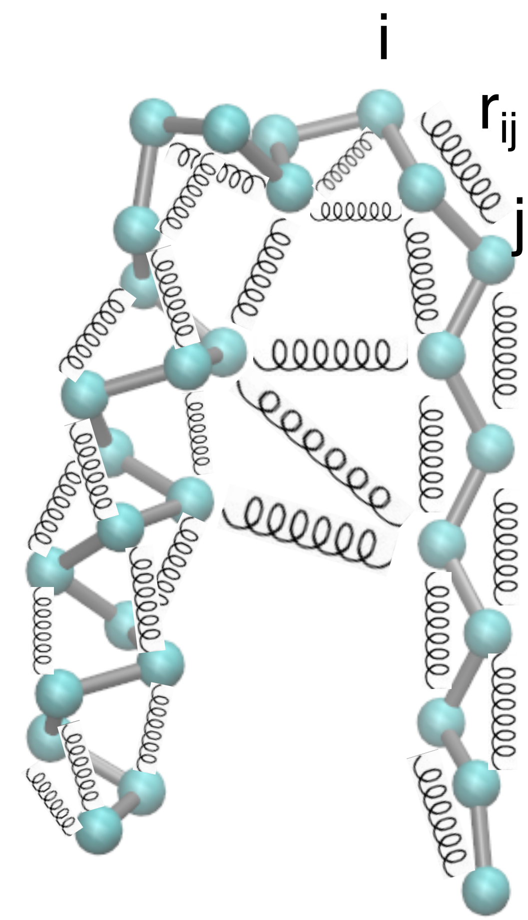
COARSE GRAINING: **NORMAL MODE ANALYSIS** (NMA)

- 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.



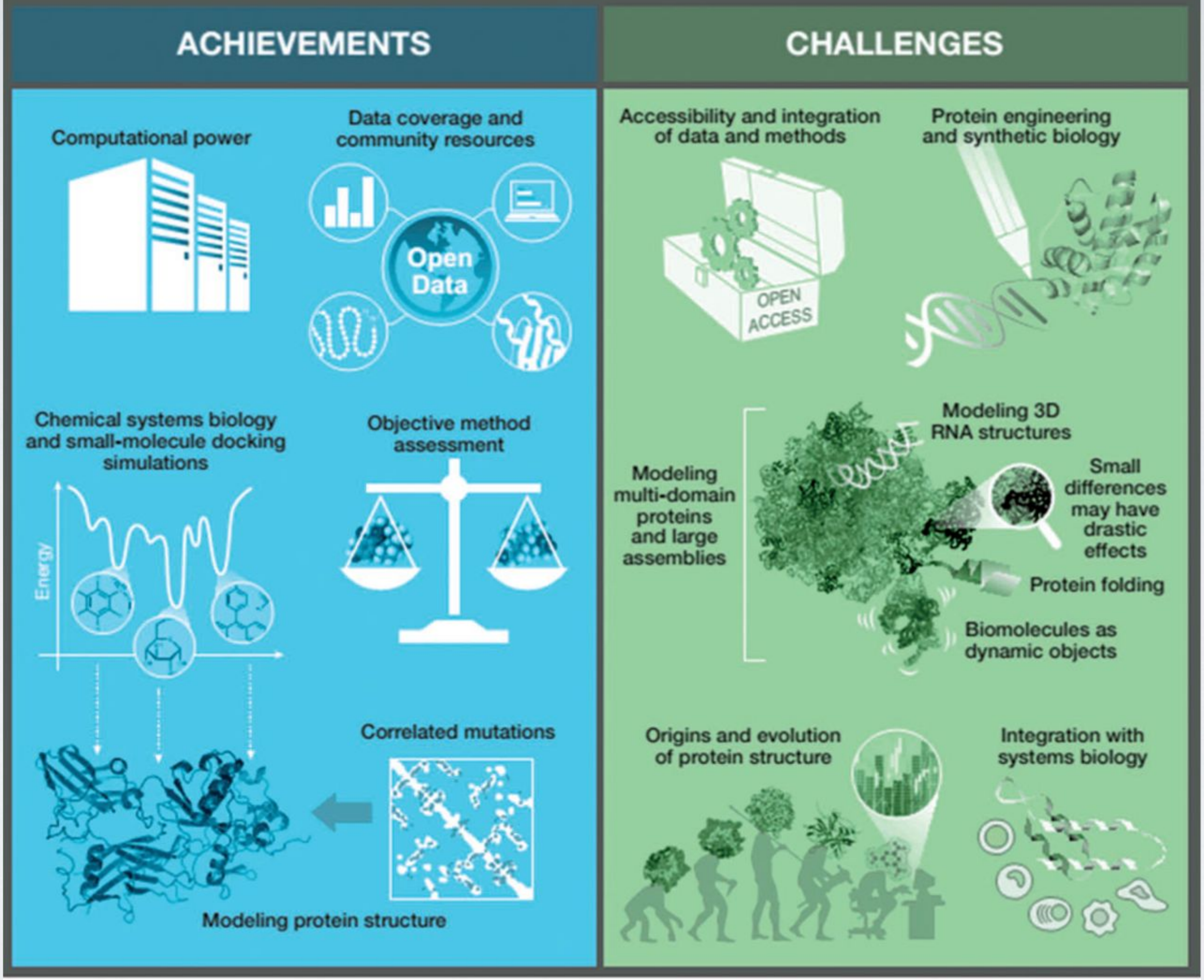
Atomistic

C. G.



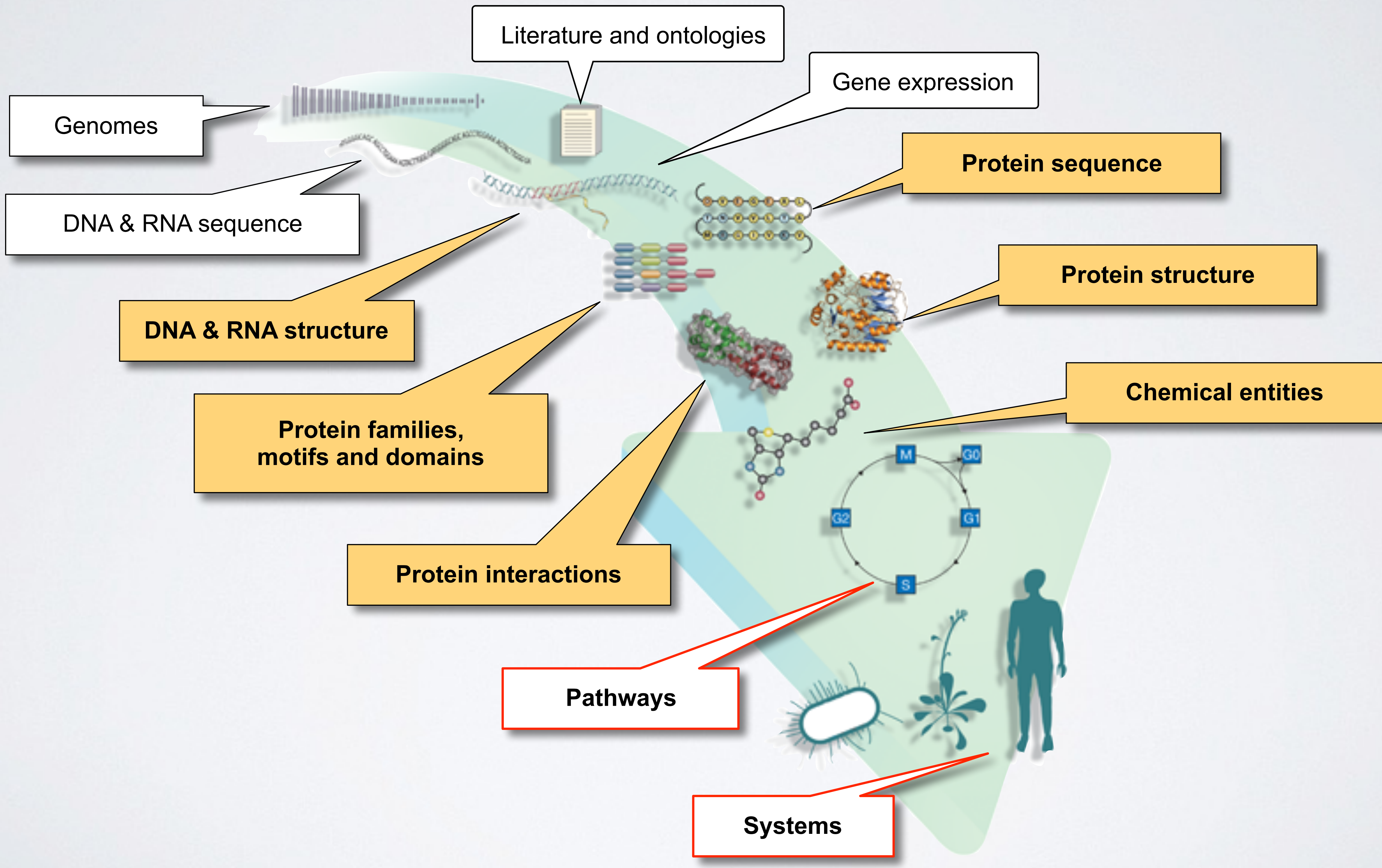
Coarse Grained

- 1 bead / 1 amino acid
- Connected by springs



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

INFORMING SYSTEMS BIOLOGY?



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.