Overview of Molecular Modeling

CHEM 430

Molecular Modeling: Introduction

What is Molecular Modeling?

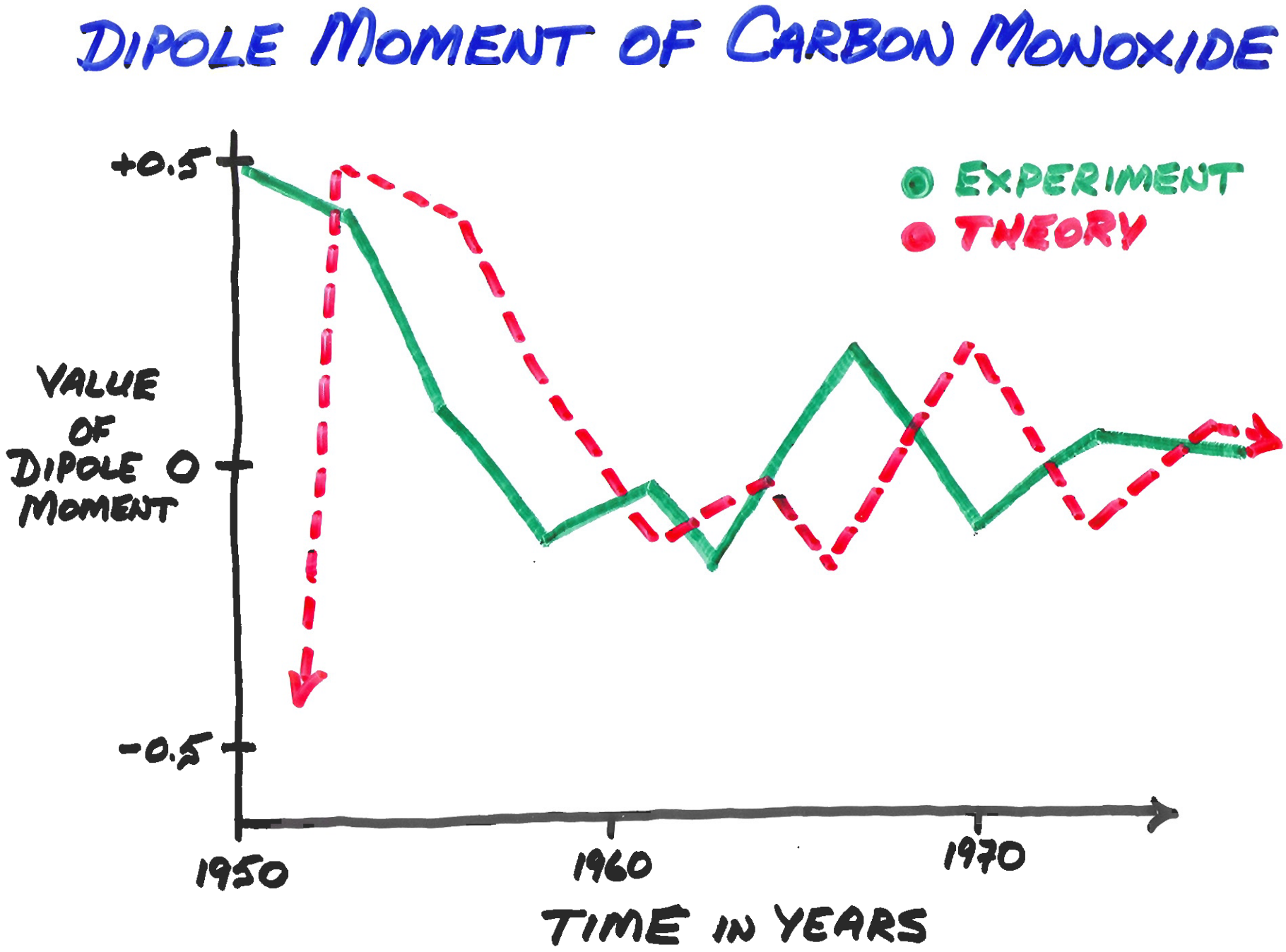
Molecular Modeling is concerned with the description of the atomic and molecular interactions that govern *microscopic* and *macroscopic* behaviors of physical systems

What is it good for?

The essence of molecular modeling resides in the connection between the *microscopic* world and the *macroscopic* world provided by the theory of statistical mechanics

Macroscopic observable (Solvation energy, affinity between two proteins, H-H distance, conformation, ...)	⟷	Average of observable over selected microscopic states

[Blatantly Stolen from Bill Goddard, Computational Chemistry GRC, circa 1986]



Experiment vs. Simulation vs. Theory

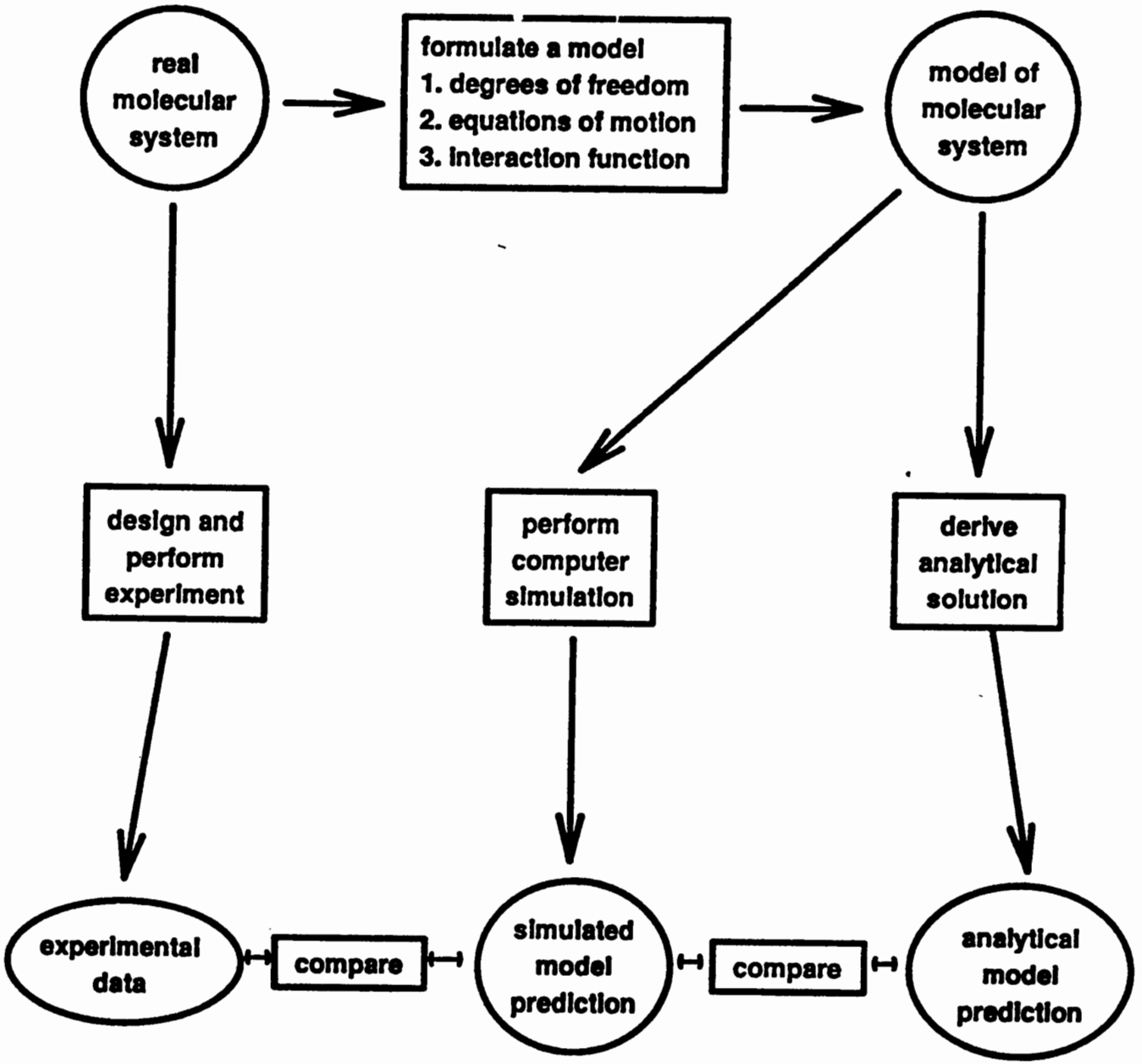


Fig. 1. Molecular models, simulation and experiment.

Types of Simulation Systems

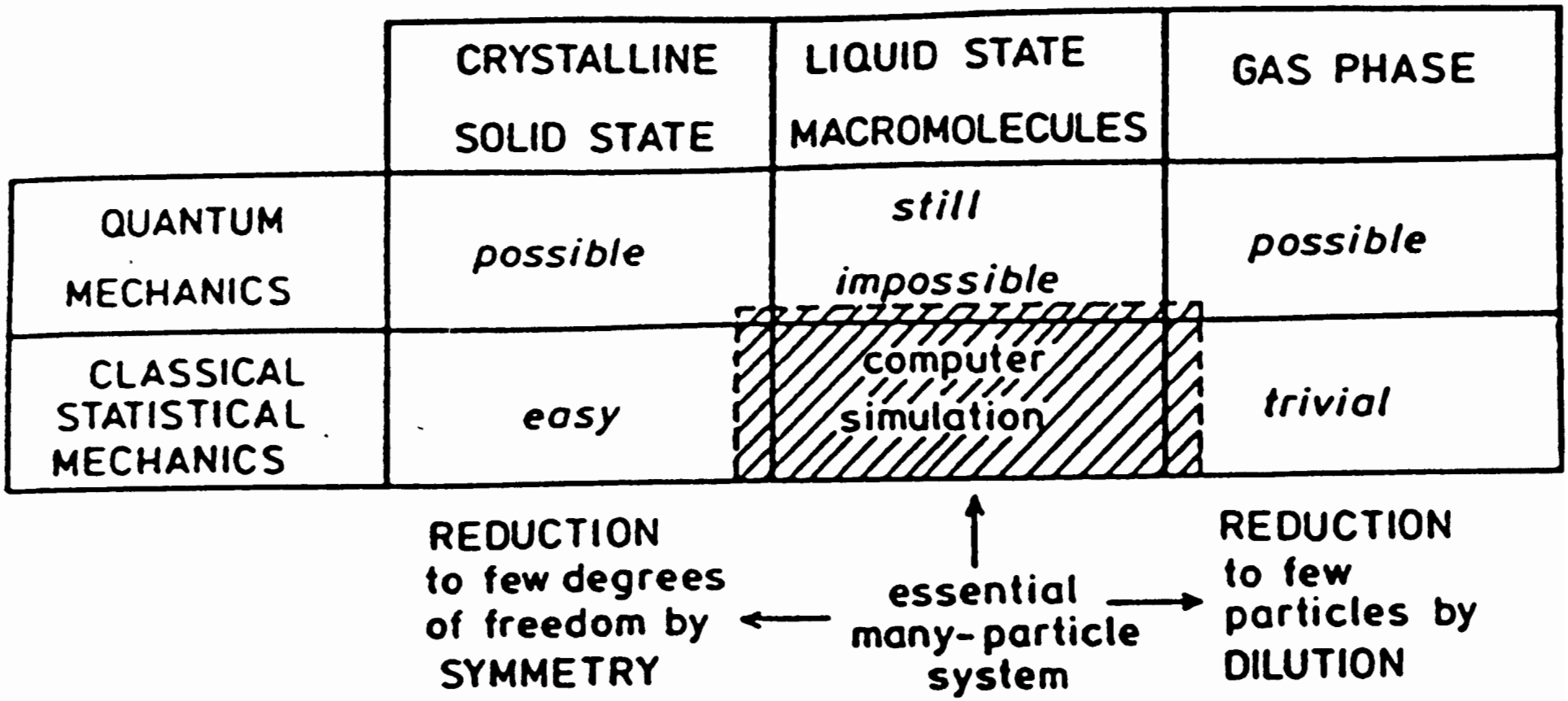


Fig. 1. Classification of molecular systems. Systems in the shaded area are amenable to treatment by computer simulation.

Model Accuracy vs. Computing Resources

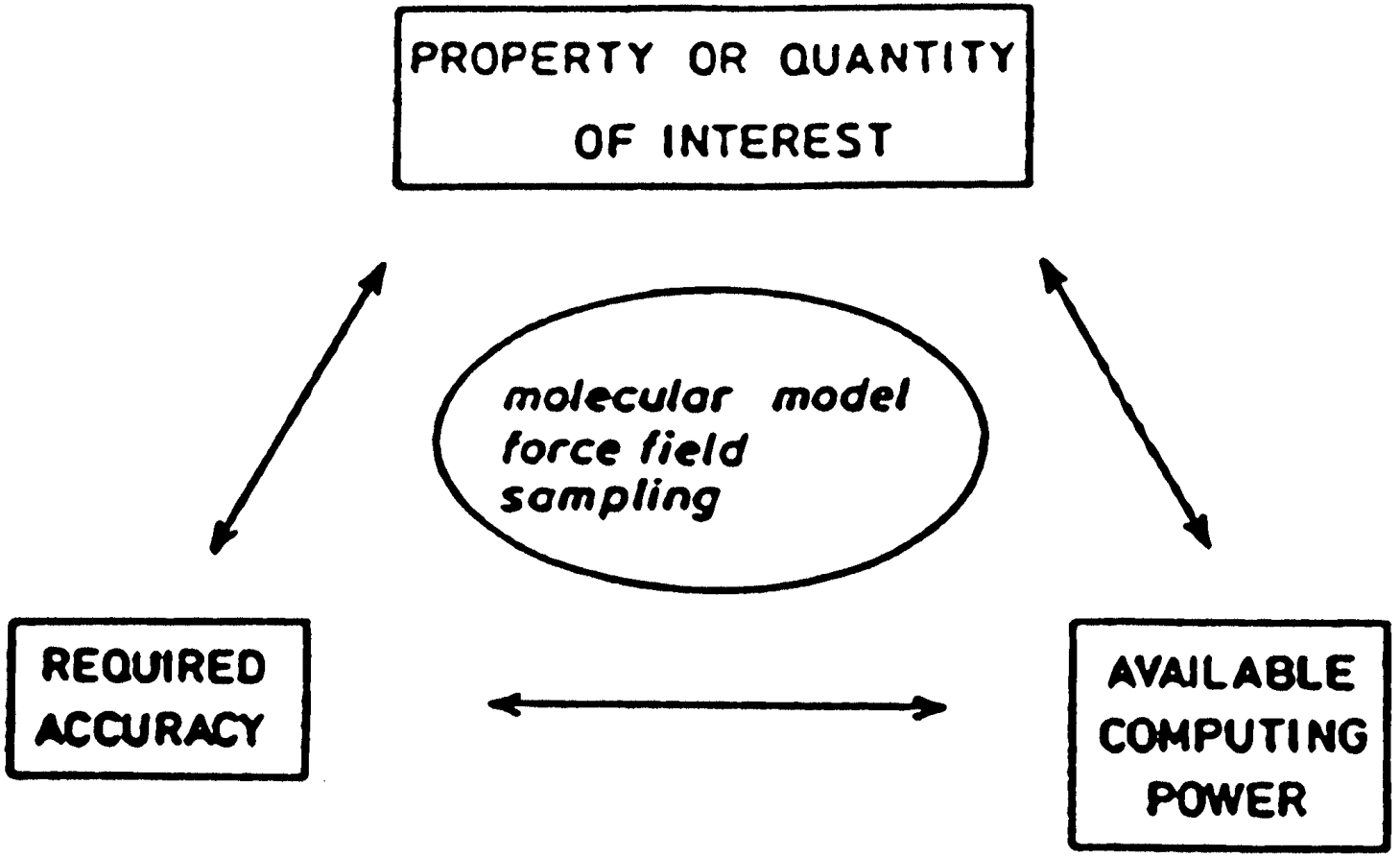


Fig. 3. Choice of molecular model, force field and sample size depends on 1) the property one is interested in (space to be searched), 2) required accuracy of the prediction, 3) the available computing power to generate the ensemble.

Molecular Motion Time Scales

Table 3.1. *Typical features of some internal motions of proteins and nucleic acids*

Motion	Spatial extent (nm)	Amplitude (nm)	Log_{10} of characteristic time (s)
Relative vibration of bonded atoms	0.2 to 0.5	0.001 to 0.01	−14 to −13
Longitudinal motions of bases in double helices (nucleic acids)	0.5	0.01	−14 to −13
Lateral motions of bases in double helices (nucleic acids)	0.5	0.1	−13 to −12
Global stretching (nucleic acids)	1 to 30	0.03 to 0.3	−13 to −11
Global twisting (nucleic acids)	1 to 30	0.1 to 1.0	−13 to −11
Elastic vibration of globular region	1 to 2	0.005 to 0.05	−12 to −11
Sugar repuckering (nucleic acids)	0.5	0.2	−12 to −9
Rotation of sidechains at surface (protein)	0.5 to 1	0.5 to 1	−11 to −10
Torsional libration of buried groups	0.5 to 1	0.05	−11 to −9
Relative motion of different globular regions (hinge bending)	1 to 2	0.1 to 0.5	−11 to −7
Global bending (nucleic acids)	10 to 100	5 to 20	−10 to −7
Rotation of medium-sized sidechains in interior (protein)	0.5	0.5	−4 to 0
Allosteric transitions	0.5 to 4	0.1 to 0.5	−5 to 0
Local denaturation	0.5 to 1	0.5 to 1	−5 to +1

Molecular Models:

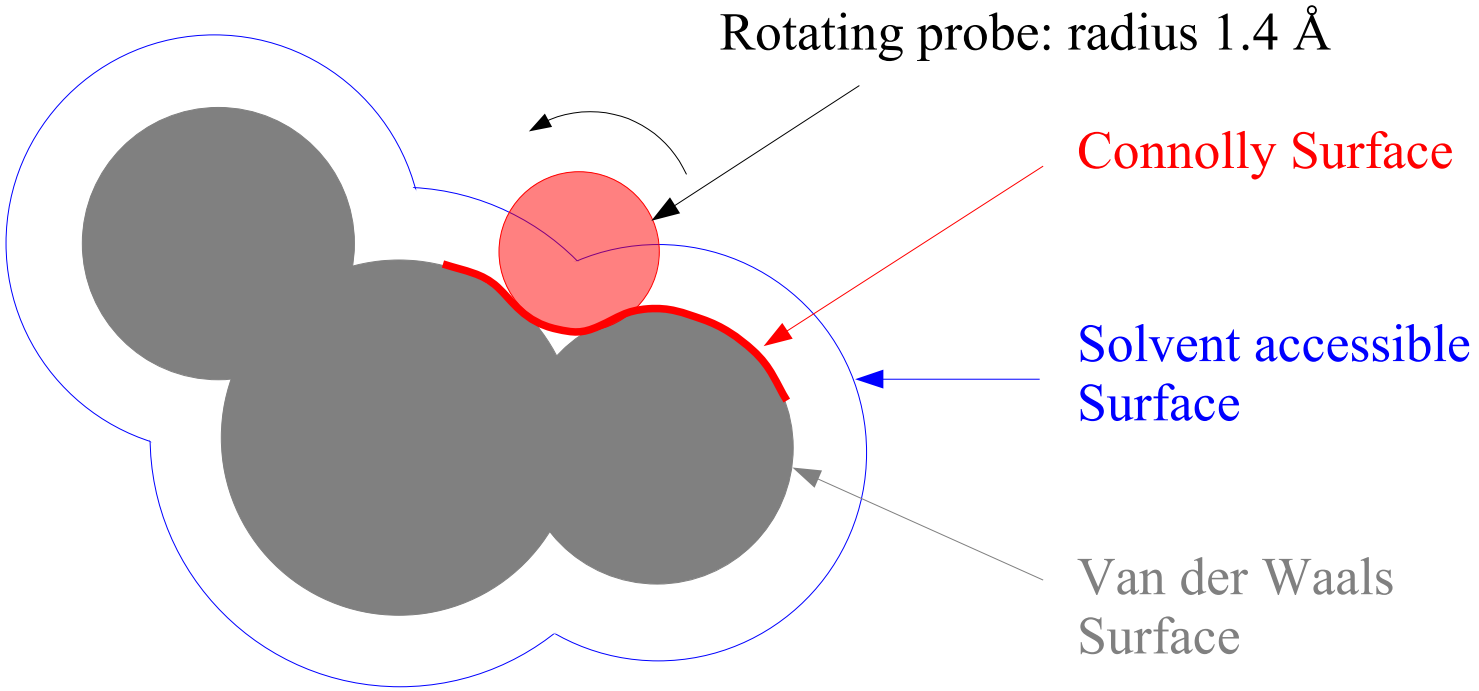
Physical

- Framework
- Space Filling

Mathematical

- Quantum Mechanics
- Classical (Empirical) Potential Functions
- Statistical or Database Derived

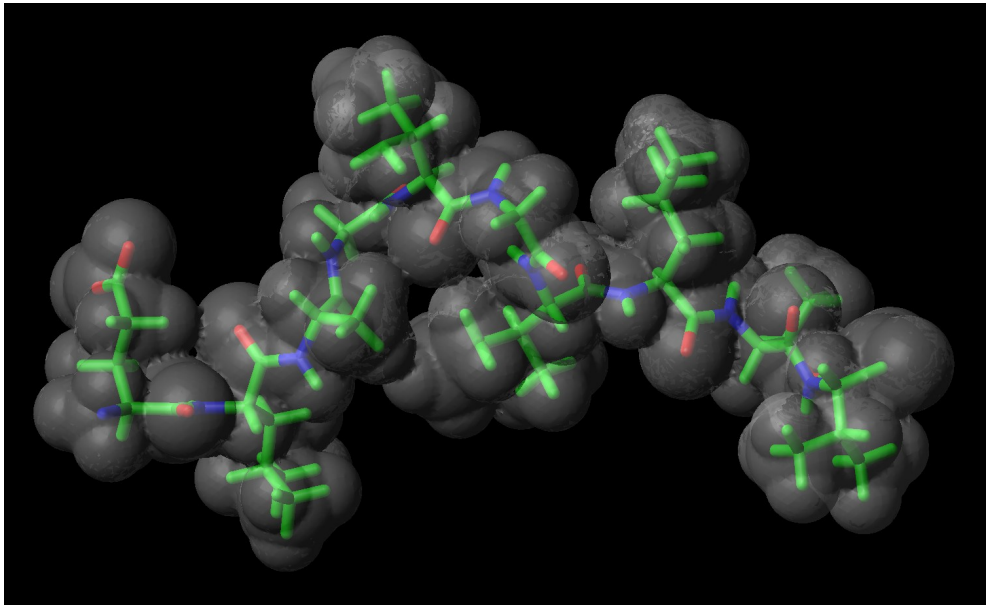
Types of Molecular Surfaces



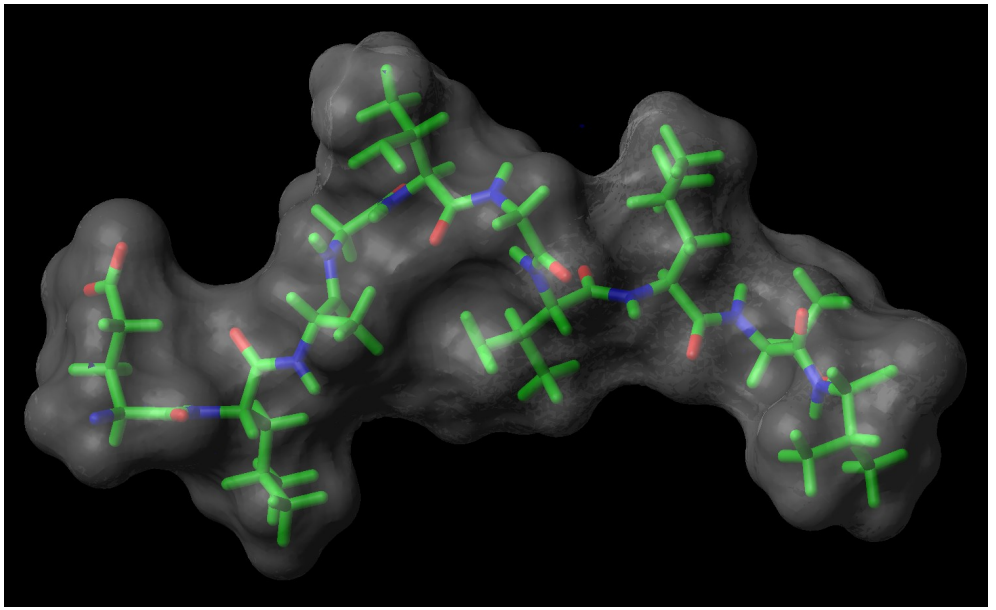
Definitions:

- Van der Waals: ensemble of van der Waals sphere centered at each atom
- Connolly: ensemble of contact points between probe and vdW spheres
- Solvent: ensemble of probe sphere centers

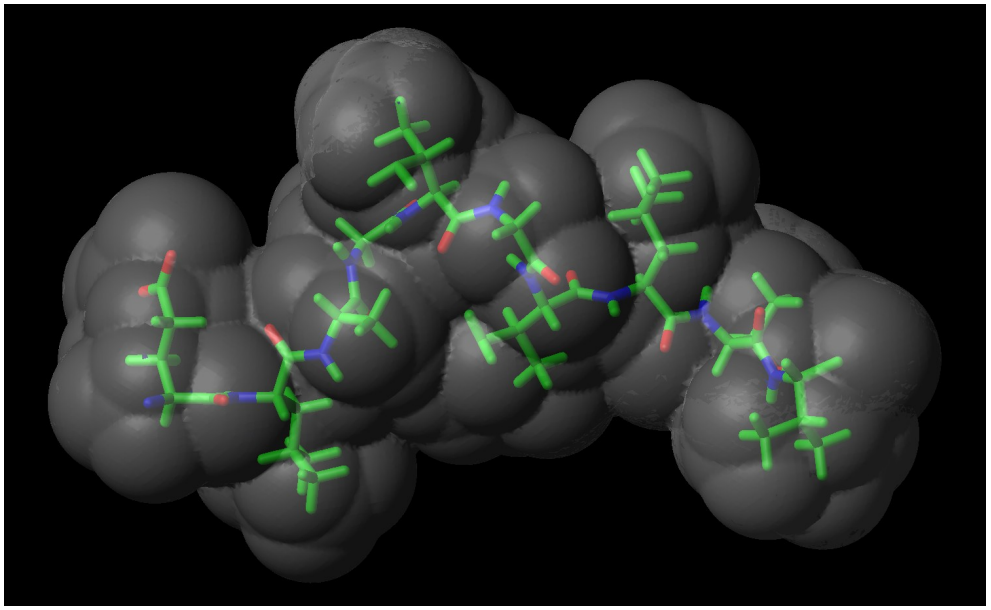
Examples of Molecular Surfaces



Van der Waals



Connolly
(Contact)

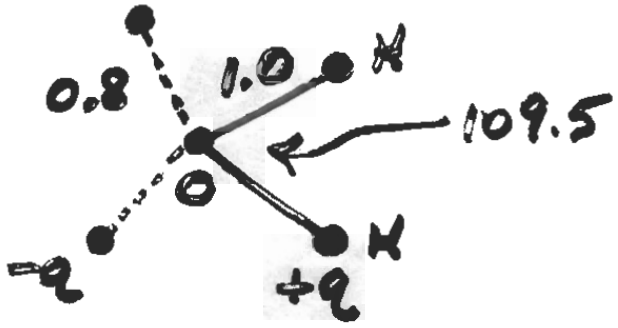


Solvent accessible

EXAMPLE: WATER

STILLINGER (ST2)

$q = 0.2357$



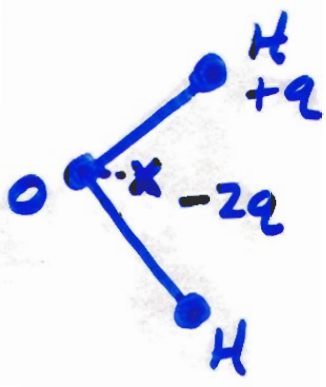
$$E_{LJ} = 0.31 \cdot (\sigma/R)^{12} - 0.31 \cdot (\sigma/R)^{6}, \quad \sigma = 3.10\,\text{Å}$$

JORGENSEN (TIP4P)

$O\text{-}X = 0.15\,\text{Å}$

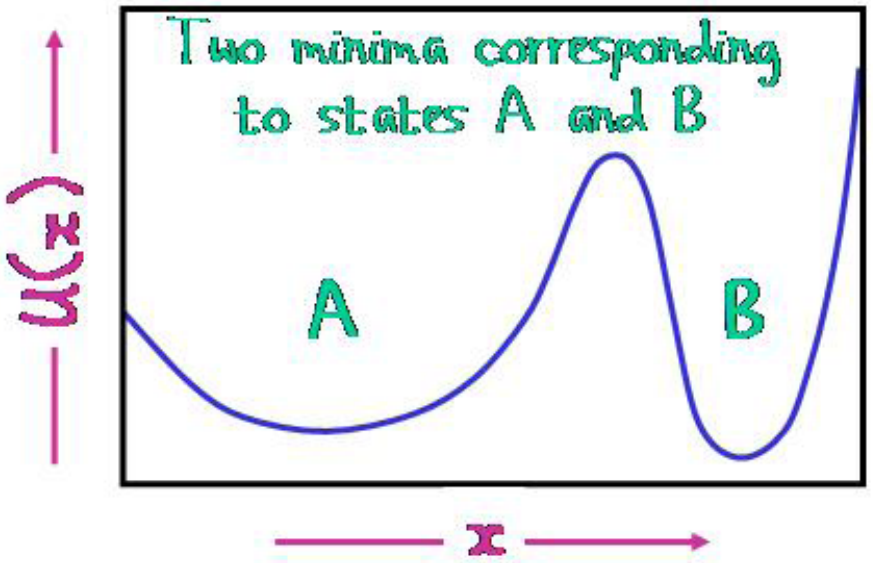
$q = 0.52$

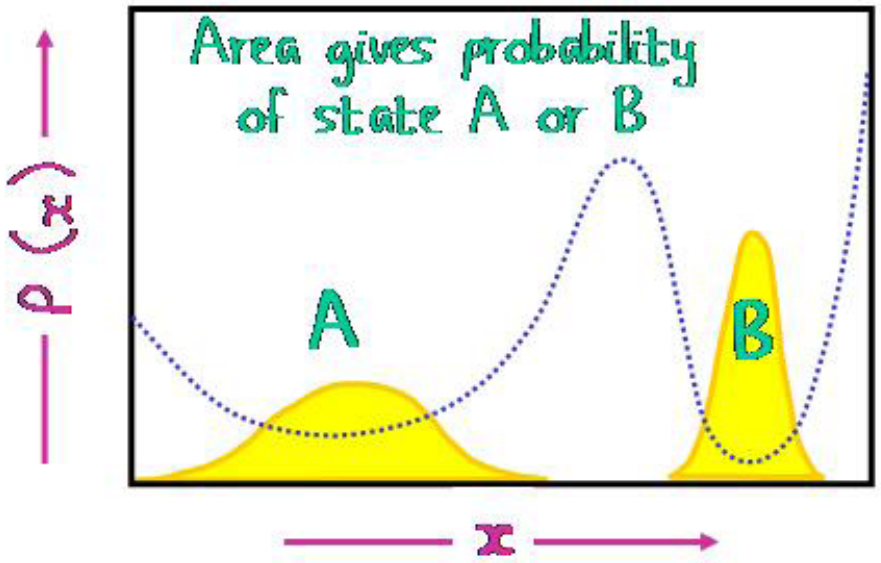
$R = 0.9572 \quad \theta = 104.52$



$$E_{LJ} = 600/R^{12} - 610/R^{6}$$

BOLTZMANN'S DISTRIBUTION





- Probability of system being at position x is

  $P(x) = \exp(-U(x)/kT) / Q$.

  $U(x)$ is Potential Energy at position x.

- Find Q, the "Partition Function", so total probability is 1.

  $Q = \sum \exp(-U(x)/kT)$

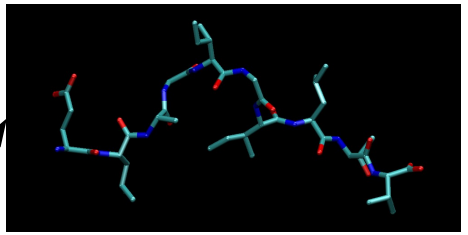
Connection between Microscopic & Macroscopic

Expectation value

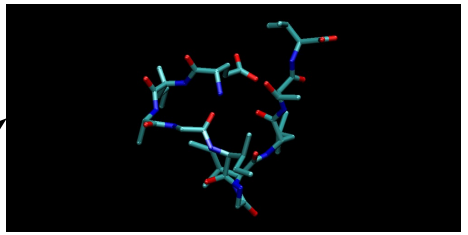
$$\langle O \rangle = \frac{1}{Z} \sum_i O_i e^{-\beta E_i}$$

Where $Z = \sum_i e^{-\beta E_i}$

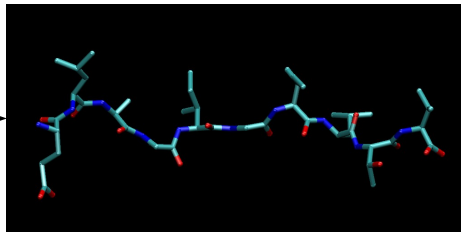
is the partition function



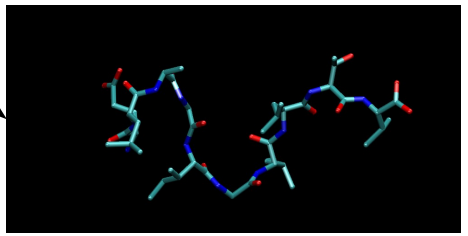
$E_1, P_1 \sim e^{-\beta E1}$



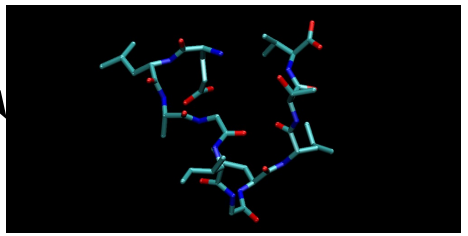
$E_2, P_2 \sim e^{-\beta E2}$



$E_3, P_3 \sim e^{-\beta E3}$



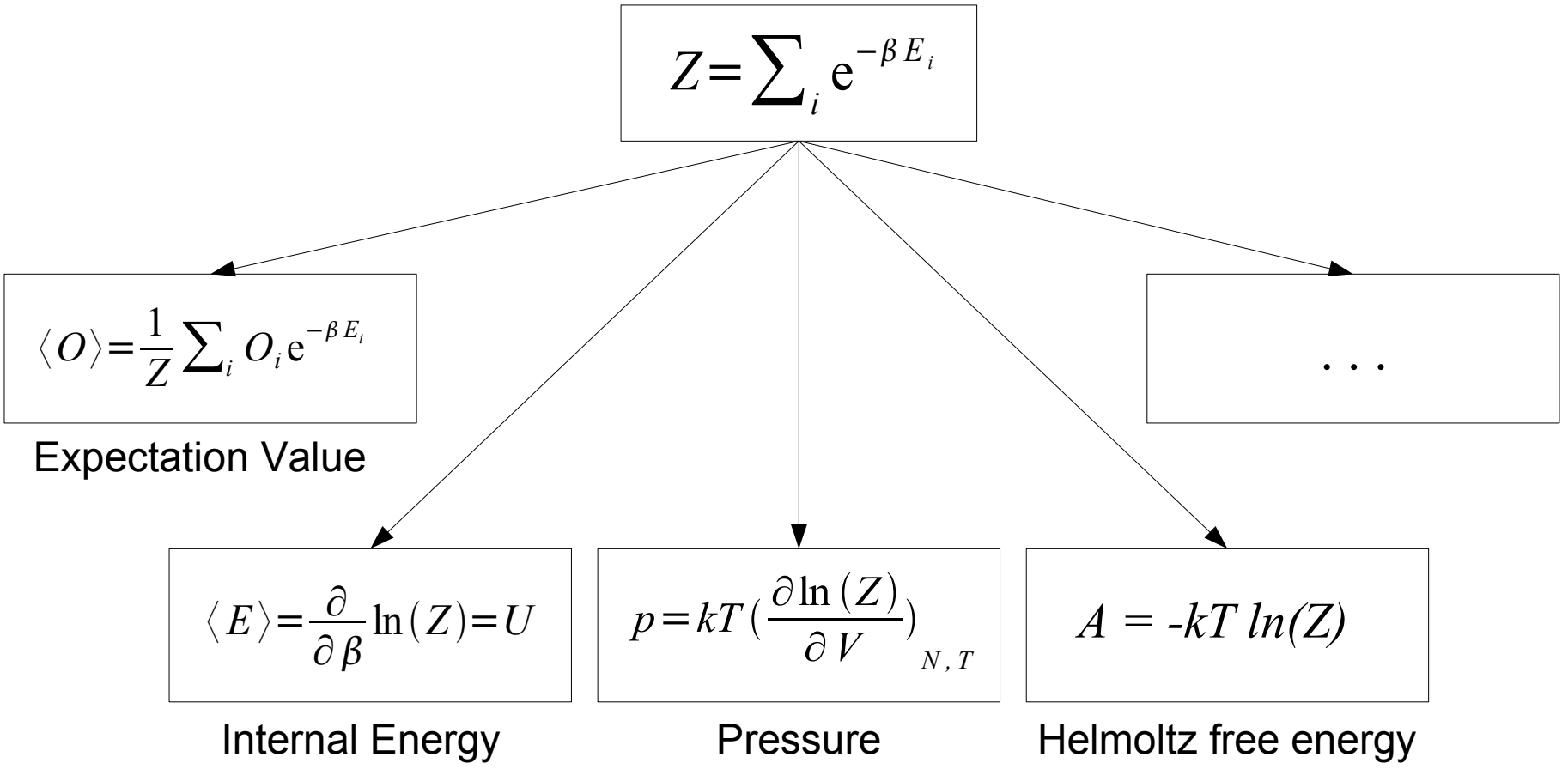
$E_4, P_4 \sim e^{-\beta E4}$



$E_5, P_5 \sim e^{-\beta E5}$

Central Role of the Partition Function

The determination of the macroscopic behavior of a system from a thermodynamical point of vue is tantamount to computing a quantity called the ***partition function***, *Z*, from which all the properties can be derived.



Computation of the Partition Function

The partition function is a very complex function to compute, and, in most cases, only numerical approximations are possible

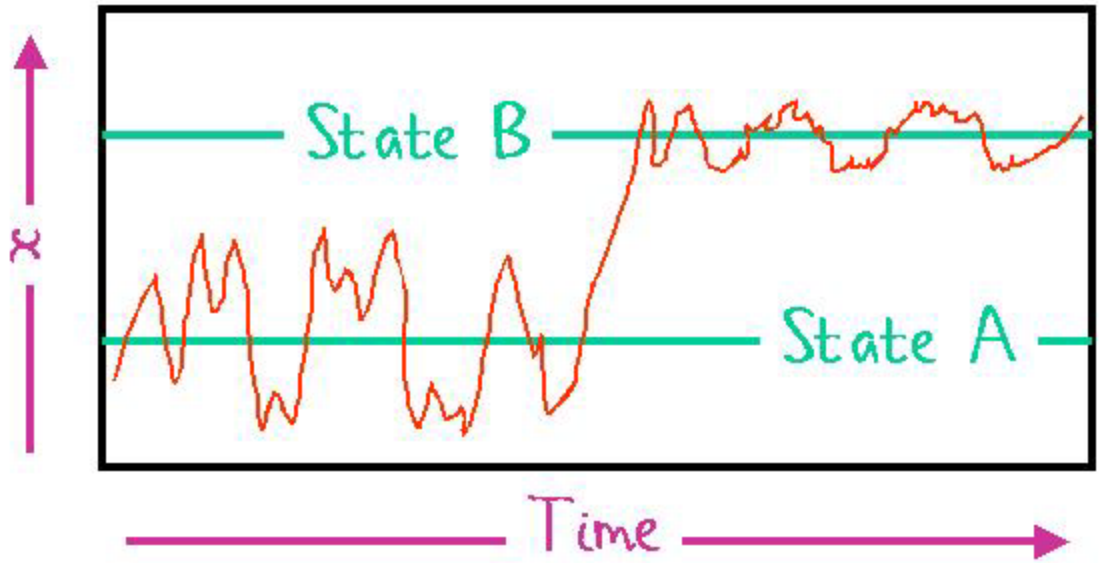
$$Z = \underbrace{\sum_i}_{2)} e^{-\beta \overbrace{E_i}^{1)}}$$

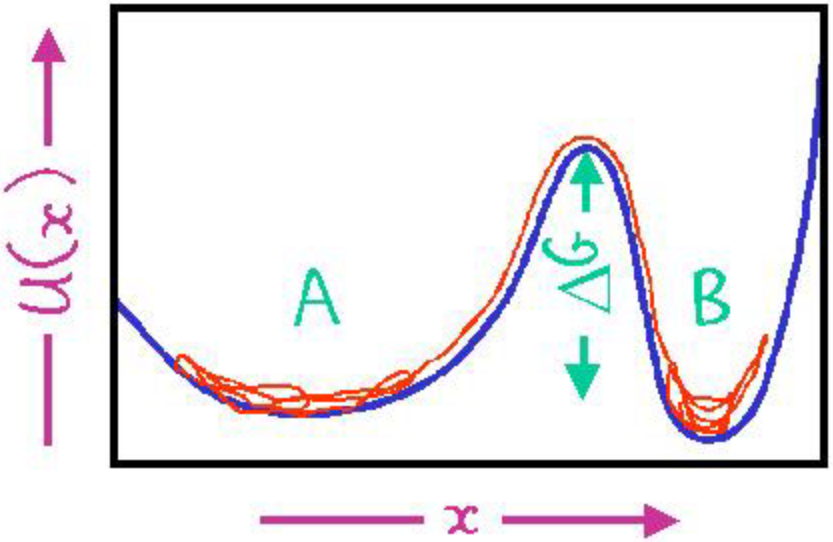
Numerical approximations require:

1) the computation of the energy of the system for microstate *i*
   - performed using semi-empirical force fields
     CHARMM / Amber / Gromacs / NAMD / Tinker ...

2) a method to sample all (or a representative portion) of the microstates accessible to the system in a given macroscopic state, i.e:
   - microcanonical sampling for fixed *N,V,E* systems
   - canonical sampling for fixed *N,V,T* systems
   - isothermic-isobaric sampling for fixed *N,P,T* systems
   - other specialized ensembles...

CROSSING ENERGY BARRIERS





- The actual transition from State A to B is very quick (a few picoseconds).
- What takes time is the waiting. The average wait before going from A to B is:

$$\tau_{A\to B} = (h/k_b T)\exp\left[+\Delta G/k_b T\right], \text{ where } \Delta G = (G_T - G_A)$$

$(h/k_b T) \sim 0.16$ picoseconds at $T = 300^\circ K$ ($27^\circ C$)

h is Planck's constant, k_b is Boltzmann's constant

Molecular Simulation: Historical Dates of Note

Theoretical milestones:

Newton (1643-1727):	Classical equations of motion: $F(t)=m\ a(t)$
Schrödinger (1887-1961):	Quantum mechanical equations of motion: $-ih\ \delta t\ \psi(t)=H(t)\ \psi(t)$
Boltzmann(1844-1906):	Foundations of statistical mechanics

Molecular dynamics milestones:

Liquids

Metropolis (1953):	First Monte Carlo (MC) simulation of a liquid (hard spheres)
Wood (1957):	First MC simulation with Lennard-Jones potential
Alder (1957):	First Molecular Dynamics (MD) simulation of a liquid (hard spheres)
Rahman (1964):	First MD simulation with Lennard-Jones potential

Proteins

Karplus (1977) & McCammon (1977)	First MD simulation of proteins
Karplus (1983):	The CHARMM general purpose FF & MD program
Kollman(1984):	The AMBER general purpose FF & MD program
Car-Parrinello(1985):	First full QM simulations
Kollmann(1986):	First QM-MM simulations

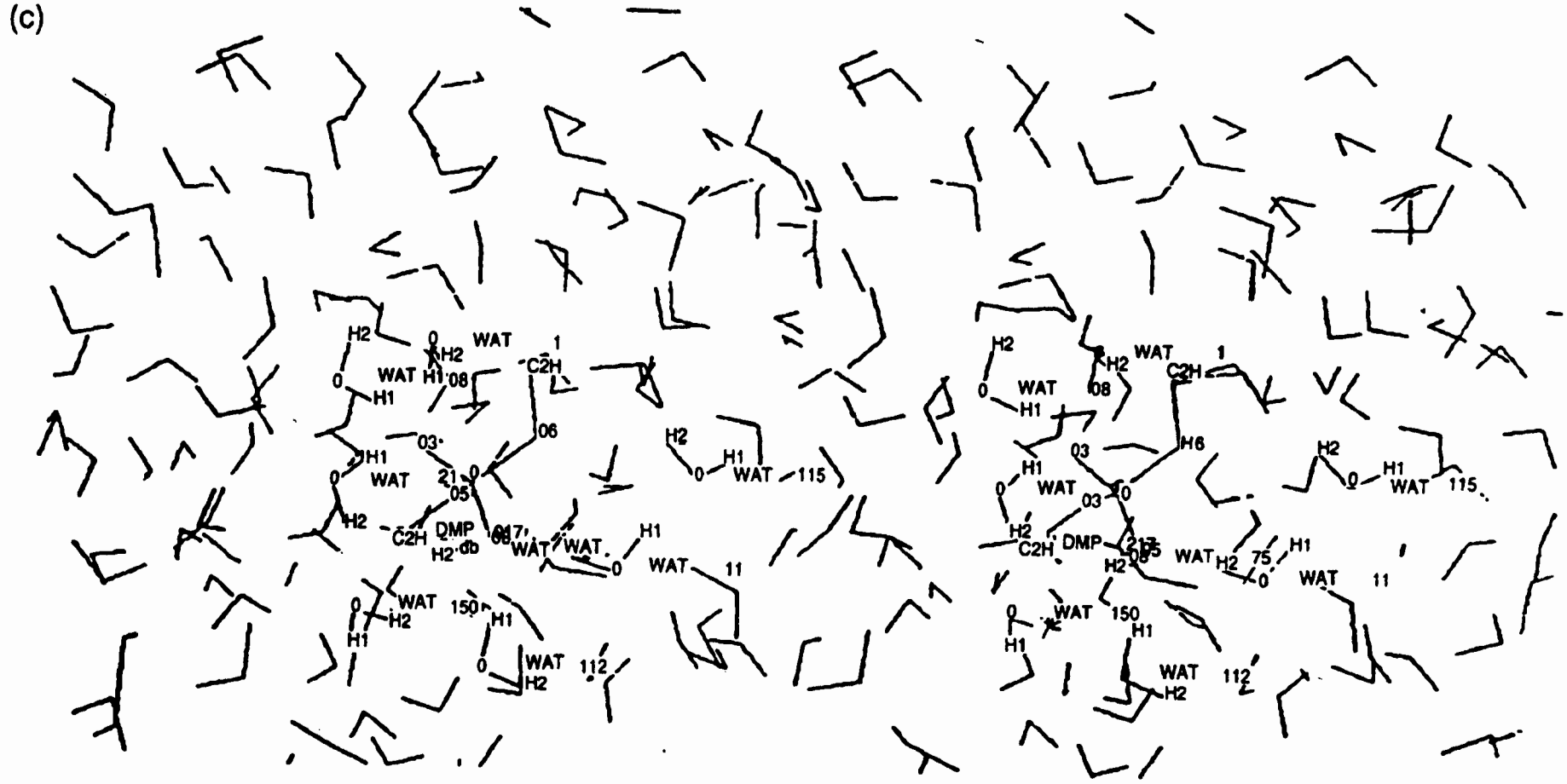


Figure 6.7 (a) Stereoscopic view of the water molecules lying near the anionic oxygens of *g,t* DMP after 5×10^5 steps (top); (b) same as (a) after 7.5×10^5 steps (centre); (c) same as (b) with a different viewpoint and all the water molecules included

Structural Characterization via RDF Analysis

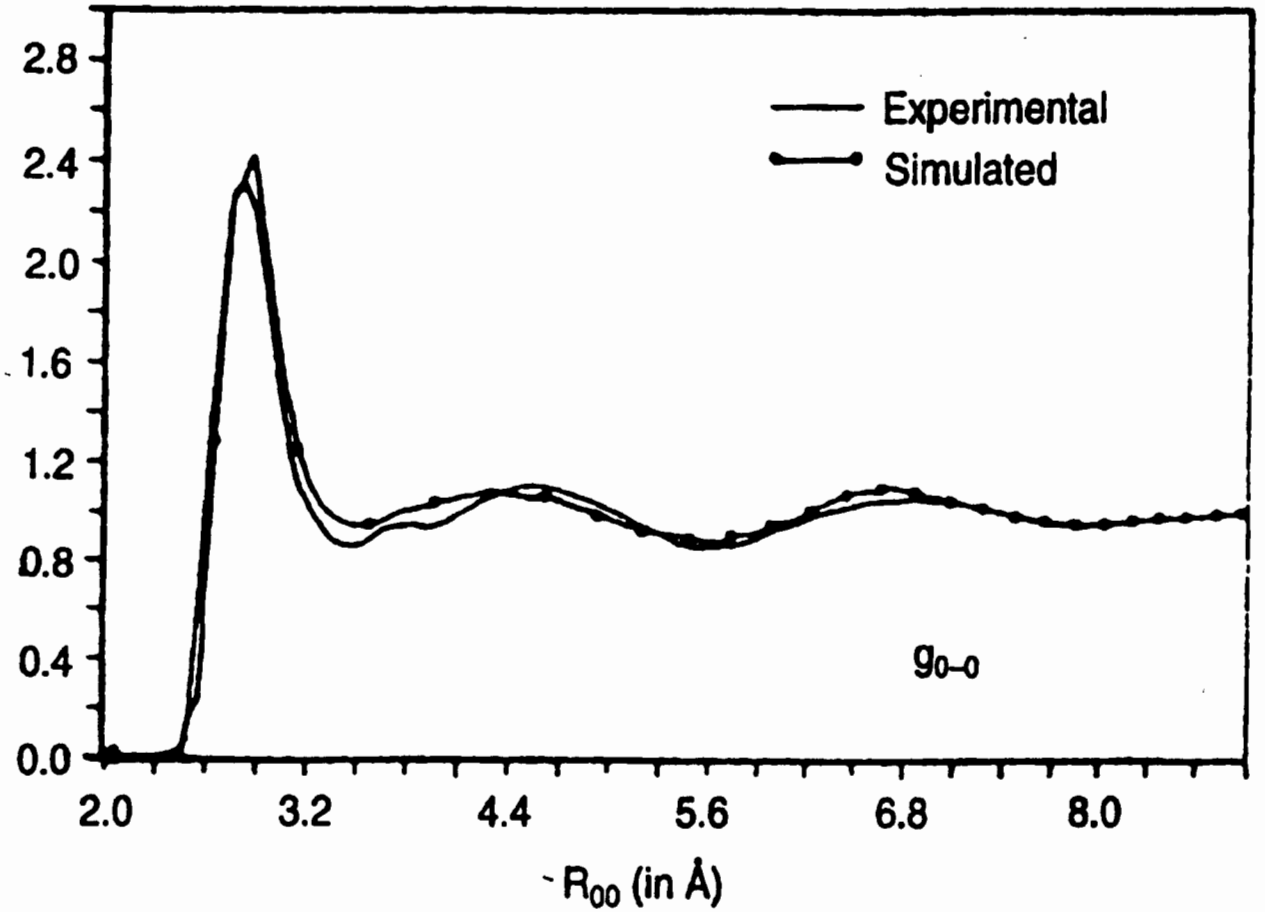


Figure 6.4 Comparison between simulated and experimental O–O radial distribution functions of liquid water (from Lie *et al.* with permission [12]).

Implicit Solvation Models

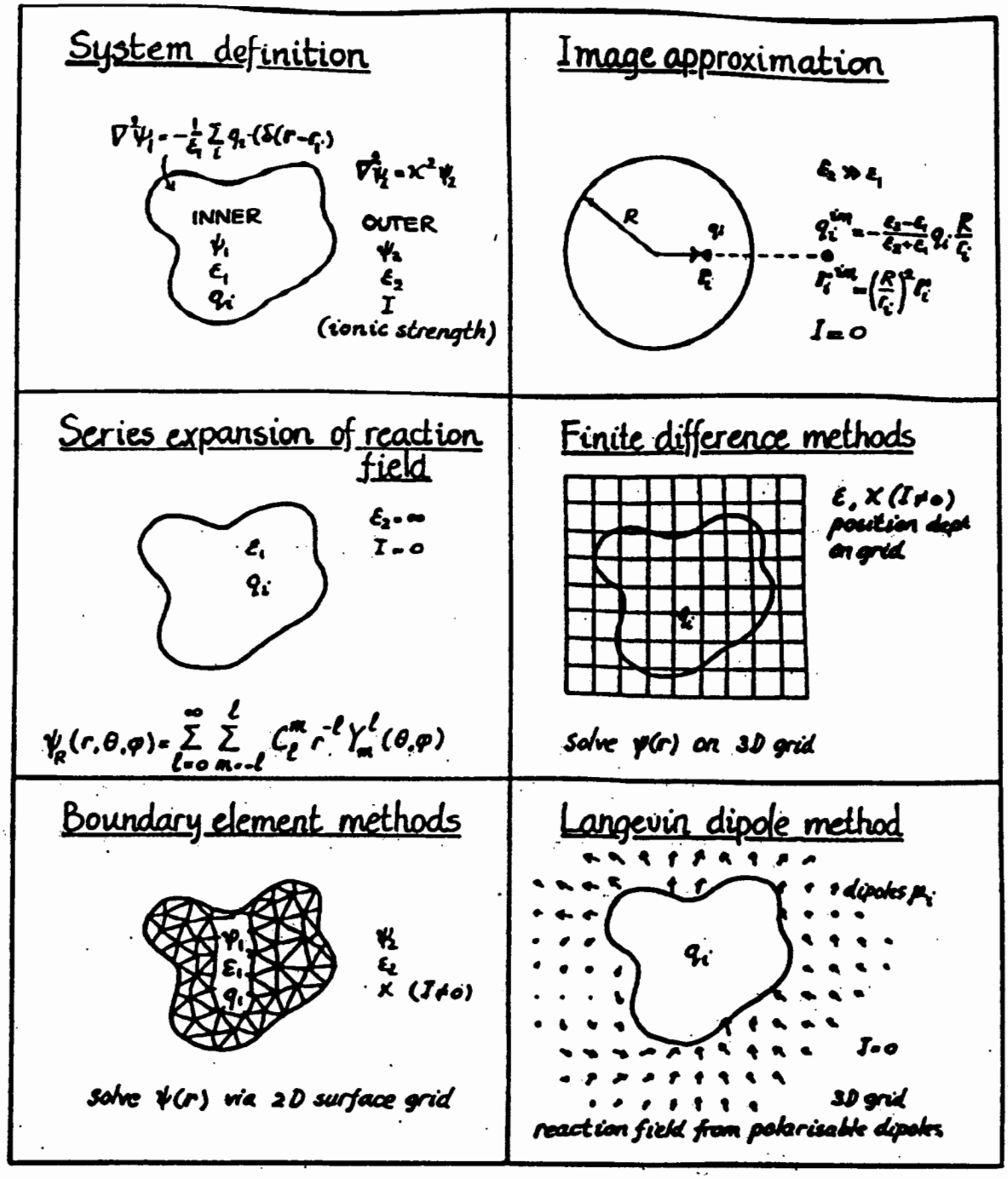
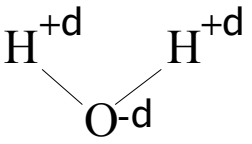
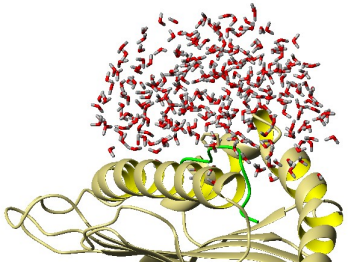


Fig. 8. Non-periodic methods for computing long-range Coulomb forces.

Treatment of the Solvent Contribution

1) Explicit water molecule model: TIP3P, ...



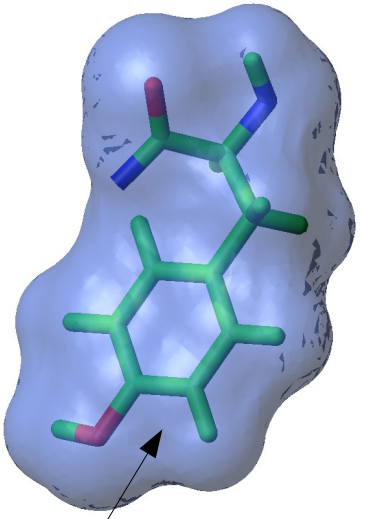


2) Implicit solvent model:

- Based on Poisson-Boltzmann Equation:

$$\nabla(\epsilon(\boldsymbol{r})\nabla\phi(\boldsymbol{r}))=\rho_{Macro}(\boldsymbol{r})+\sum_i q_i n_i^0 \exp(-\beta q_i \phi(\boldsymbol{r}))$$

- or an approximation...
  - -ACE potential (Schaeffer & al.)
  - -SASA potential (Caflish & al.)
  - -EEF1 potential (Lazaridis & al.)



For a discussion of theoretical aspects of implicit solvent models, see Roux & Simonson (***Biophys. Chem.*** 1999, 78:1-20)

Double Dynamic Programming Alignment Algorithms

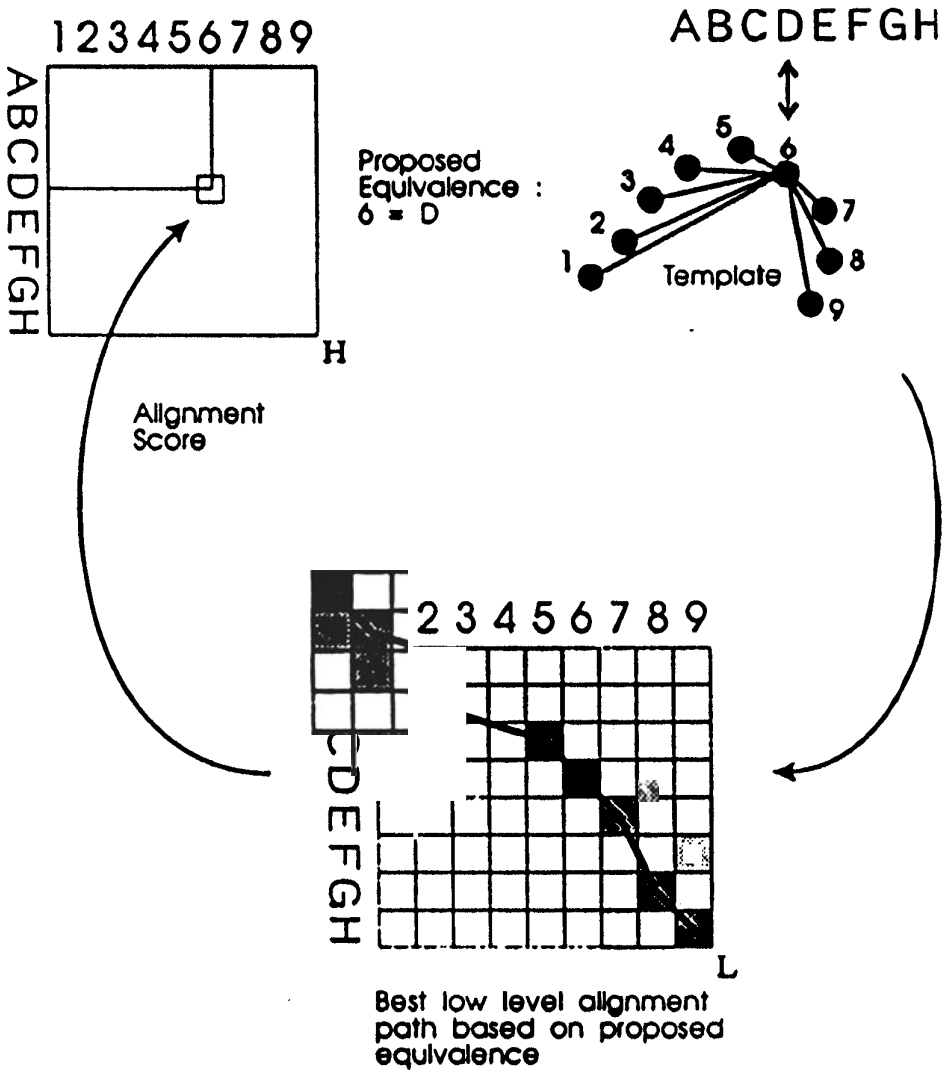
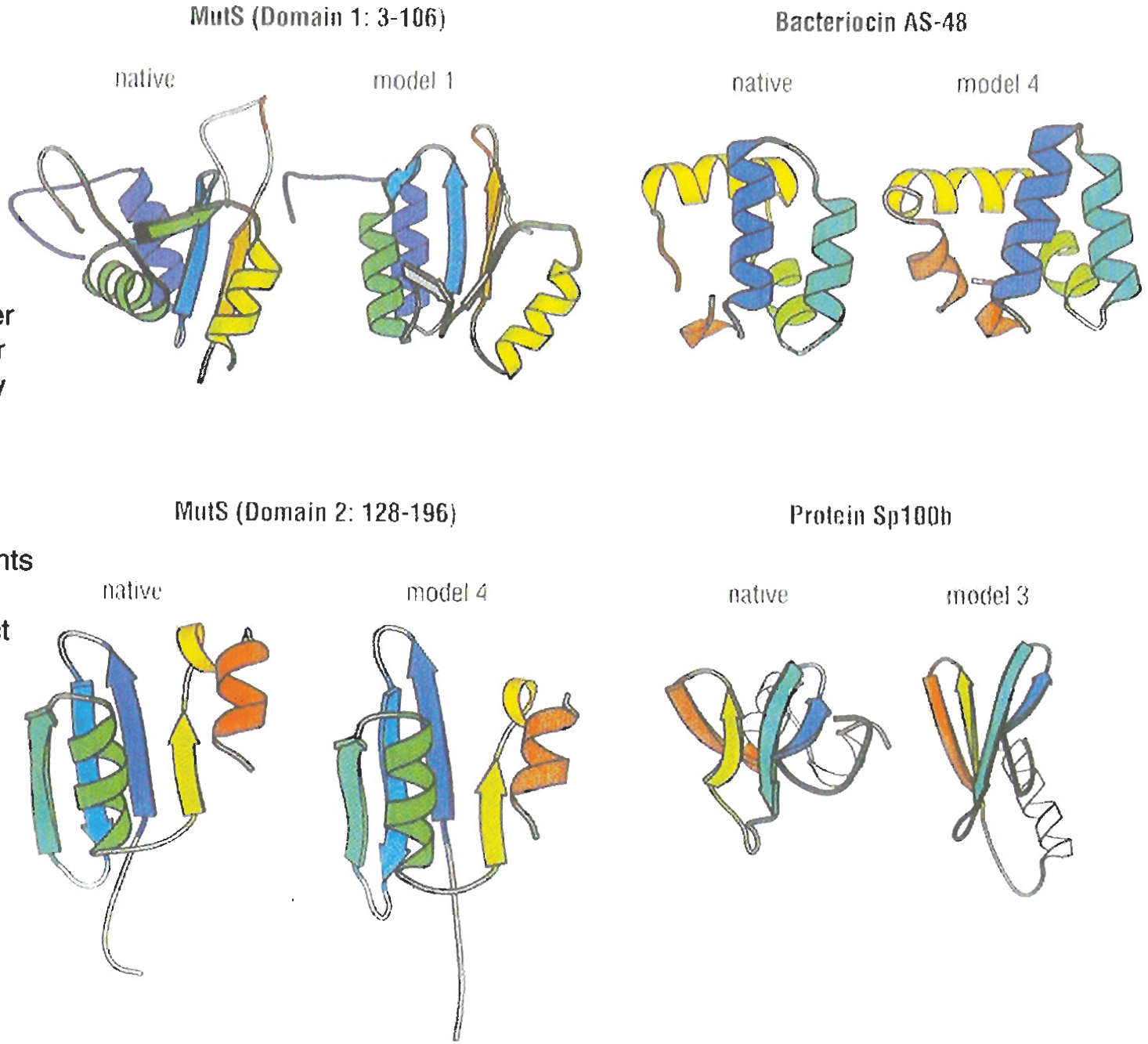


Fig. 4. An illustration of the Double Dynamic Programming algorithm applied to the sequence threading problem. A sequence of amino acids A–H (one letter code) is being threaded onto coordinate positions 1–9 of a structural template. Given the proposed equivalence that residue D lies on position 6, a matrix L of the scores of all other equivalences can be constructed. A best path (or alignment) is then found through this by application of the standard Dynamic Programming algorithm. The overall score for this alignment (which is an indication of how well residue D fits on position 6) is recorded in the matrix H. All potential equivalences are evaluated in this way (filling the matrix H with values) and the best consistent selection of these is found by application of the Dynamic Programming algorithm to the H matrix. This double application of the alignment algorithm at two levels gives rise to the name.

Rosetta Uses Fragment Library + Monte Carlo Search

Examples of the best-center cluster found by *Rosetta* for some test proteins. In many cases the overall fold is predicted well enough to be recognizable. However, relative positions of the secondary structure elements are almost always shifted somewhat from their correct values.



AlphaFold2 Structure Predictions from CASP14