Structural Bioinformatics

Elodie Laine

Master BIM-BMC Semestre 3, 2014-2015

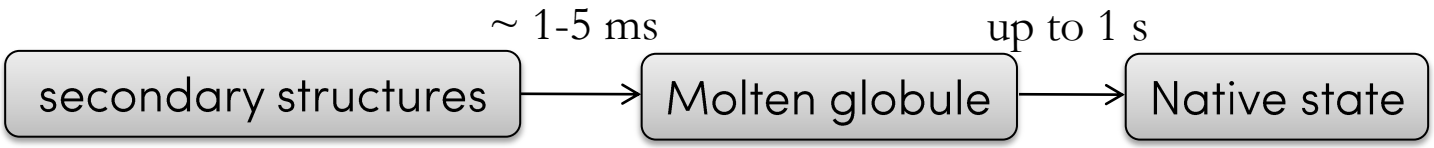
Laboratoire de Biologie Computationnelle et Quantitative (LCQB)
e-documents: http://www.lcqb.upmc.fr/laine/STRUCT
e-mail: elodie.laine@upmc.fr

Lecture 2 – Energy Functions

Protein folding

Protein folding is the process by which a polypeptide chain acquires its correct three-dimensional structure to achieve the biologically native state.

Small proteins can fold spontaneously while larger ones require the assistance of **chaperones**.

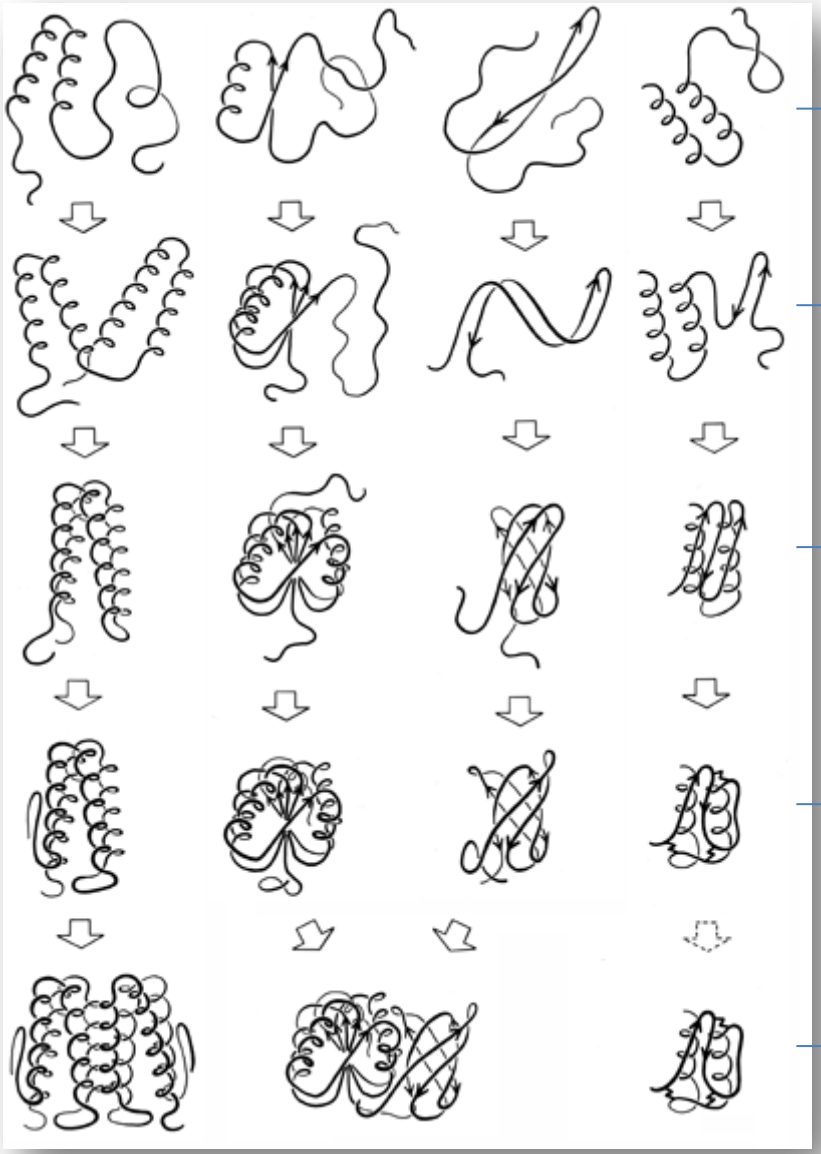


The number of possible conformations is huge and their systematic exploration is unrealistic. The native structure generally corresponds to one of these conformations: **the free energy minimum.**

Time necessary for a protein to adopt its native structure: between 1 ms and 1 s.

Protein folding

speculative general scheme of protein folding for each of the major structure categories



Nucleation

Growth and coalescence

to form

regular secondary structure

readjustment for max overall stability

quaternary association

The thermodynamic hypothesis

The Nobel Prize in Chemistry 1972

"for his work on ribonuclease, especially concerning the connection between the amino acid sequence and the biologically active conformation"

"for their contribution to the understanding of the connection between chemical structure and catalytic activity of the active centre of the ribonuclease molecule"

Christian B. Anfinsen

1/2 of the prize

USA

National Institutes of Health
Bethesda, MD, USA

b. 1916
d. 1995

Stanford Moore

1/4 of the prize

USA

Rockefeller University
New York, NY, USA

b. 1913
d. 1982

William H. Stein

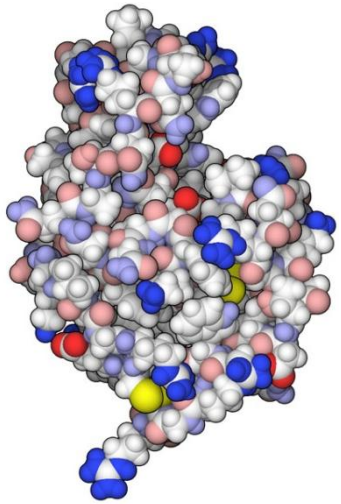
1/4 of the prize

USA

Rockefeller University
New York, NY, USA

b. 1911
d. 1980

The native structure of a protein corresponds to **minimum of free energy**



Globular proteins are only **marginally stable**. The free energy difference between the native state and the ensemble of denatured conformations is **5-15 kcal/mol**.
(H-bond: 2-5kcal/mol)

Free energy

The protein in solution is viewed as a **statistical ensemble**

$$\Delta G = \Delta H - T\Delta S$$

Gibbs or Helmholtz free energy

Energy available for thermodynamic work

~5-15 kcal/mol

Enthalpy or internal energy

Internal interactions

up to ~100 kcal/mol

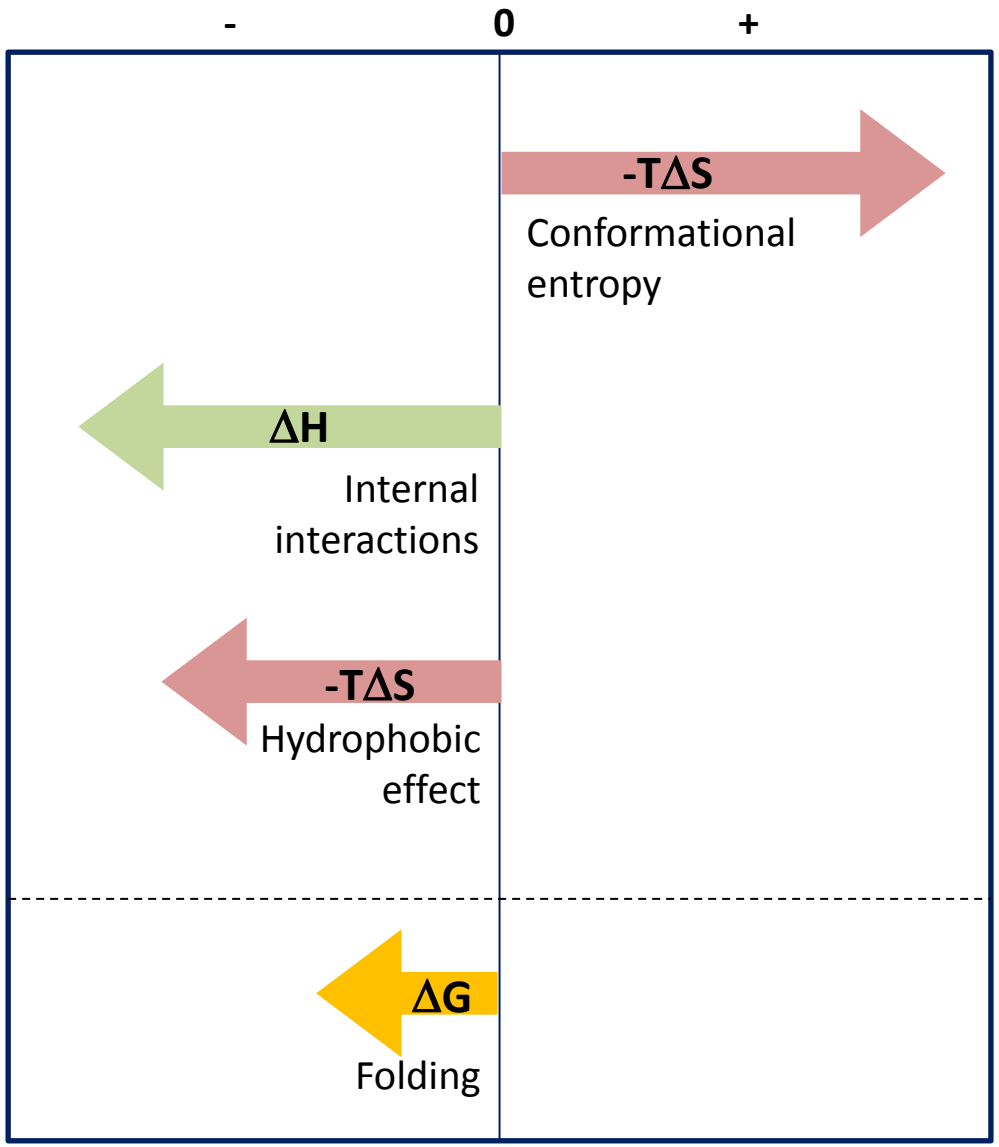
Entropy by temperature

Hydrophobic effect and conformational entropy

up to ~100 kcal/mol

Free energy

Favorable free energy of folding is a net result of thermodynamic forces

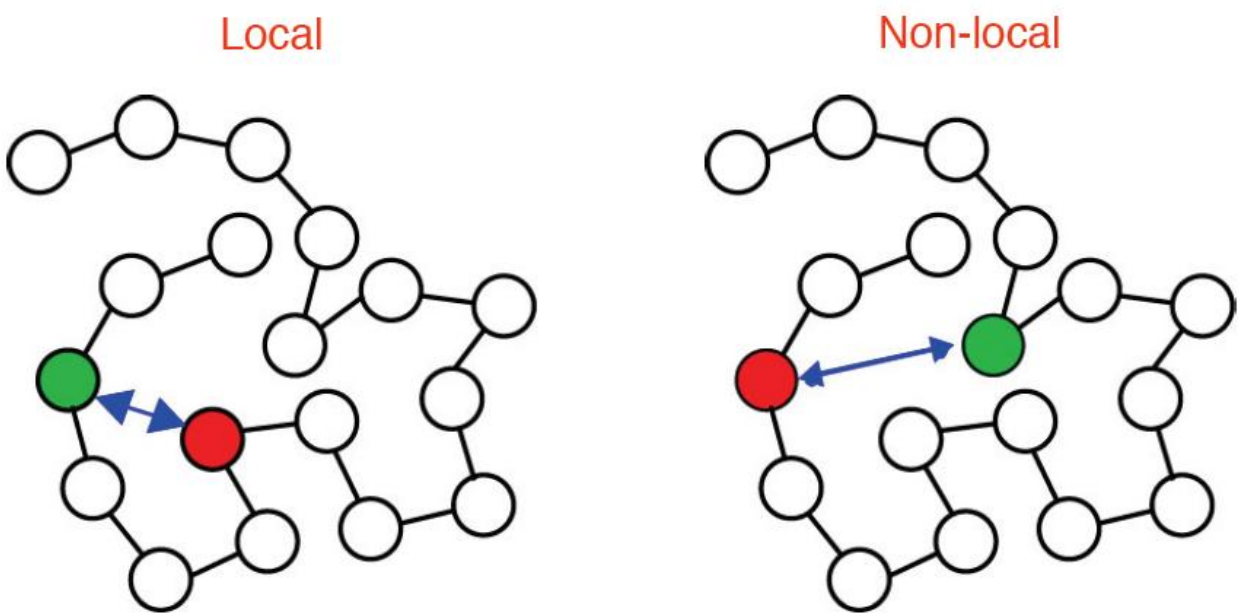


Interatomic interactions

Amino acids of a protein are joined by **covalent bonding interactions** (primary structure).

The 3D fold is stabilized by **non-bonding interactions** (tertiary structure):

- Electrostatic interactions ~5 kcal/mol
- Hydrogen-bond interactions ~3-7 kcal/mol
- Van Der Waals interactions ~1 kcal/mol
- Hydrophobic interactions < 10 kcal/mol



Energy functions

Semi-empirical potentials

- analytical forms describing interactions which parameters are fitted to:
  - experimental data
  - quantum mechanics calculations

Statistical potentials

- analytical forms describing interactions which parameters are derived from a database of known structures
- sequence-structure association frequencies converted to free energies

Energy functions

Semi-empirical potentials

- **analytical forms describing interactions which parameters are fitted to:**
  - **experimental data**
  - **quantum mechanics calculations**

Statistical potentials

- analytical forms describing interactions which parameters are derived from a database of known structures
- sequence-structure association frequencies converted to free energies

Semi-empirical potentials

These potentials are **analytical expressions** that describe inter-atomic interactions. They represent **molecular mechanics** models of proteins containing:

- some chosen interactions
- a chosen functional that describes and links them

Schrödinger **Newton**

$$i\hbar\frac{\partial}{\partial t}\Psi = H\Psi$$

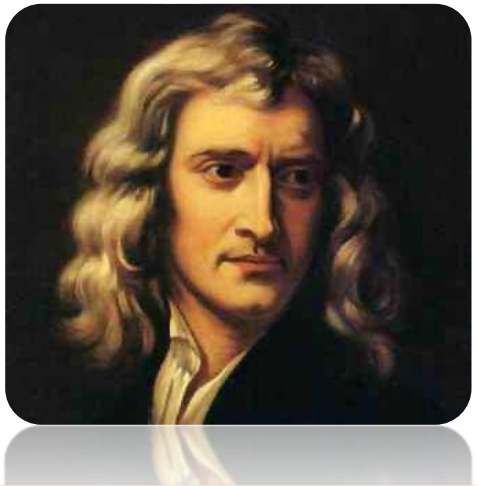
Born-Oppenheimer

Additivity

Transferability

$$M\ddot{\mathbf{x}}(t) = F(\mathbf{x}(t)) = -\nabla V(\mathbf{x}(t))$$

Empirical



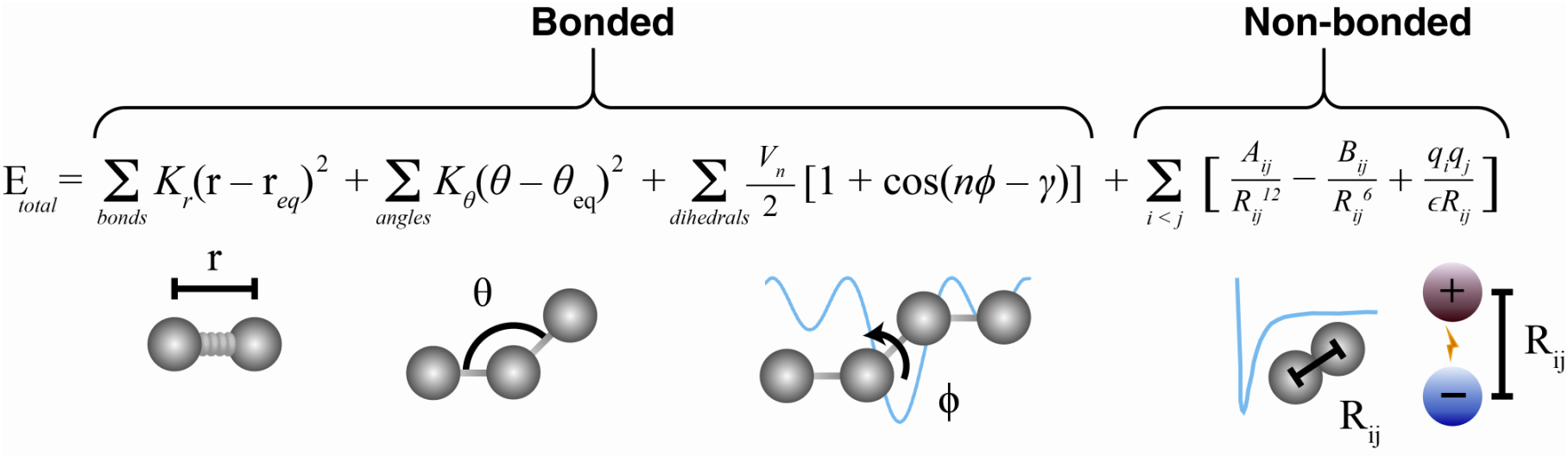
Their general form is:

$$\mathbf{E} = \mathbf{E}_{bond} + \mathbf{E}_{angle} + \mathbf{E}_{torsion} + \mathbf{E}_{non\text{-}bonded} + \mathbf{E}_{others}$$

Molecular mechanics energy

- An example: AMBER force field

$$E_{total} = \underbrace{\sum_{bonds} K_r(r - r_{eq})^2 + \sum_{angles} K_\theta(\theta - \theta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)]}_{\text{Bonded}} + \underbrace{\sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}} \right]}_{\text{Non-bonded}}$$



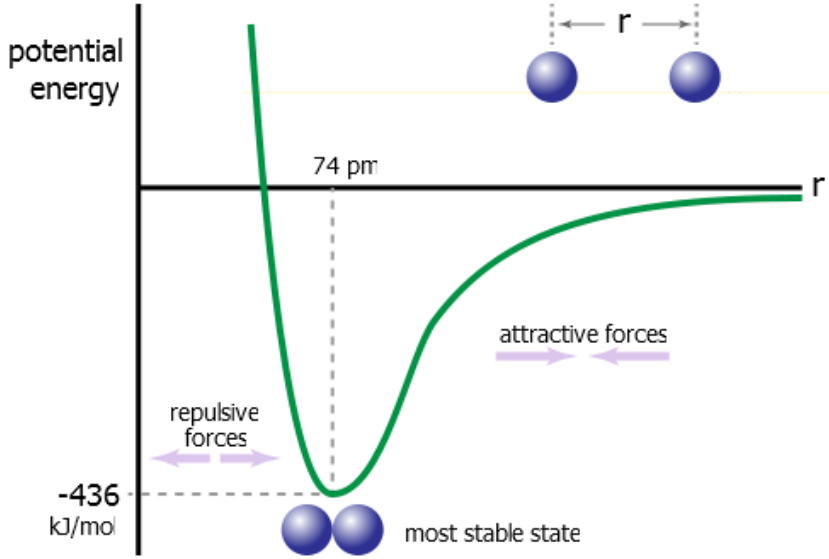
O(N)

O(N₂) ⟶ *O(NlogN)*

Many more… CHarMM, OPLS…

Bonded interactions

How can we represent the variation of the energy corresponding to a covalent bond stretching and bending?



Covalent bond potential energy: which function to model this curve ?

Bonded interactions

- Morse potential

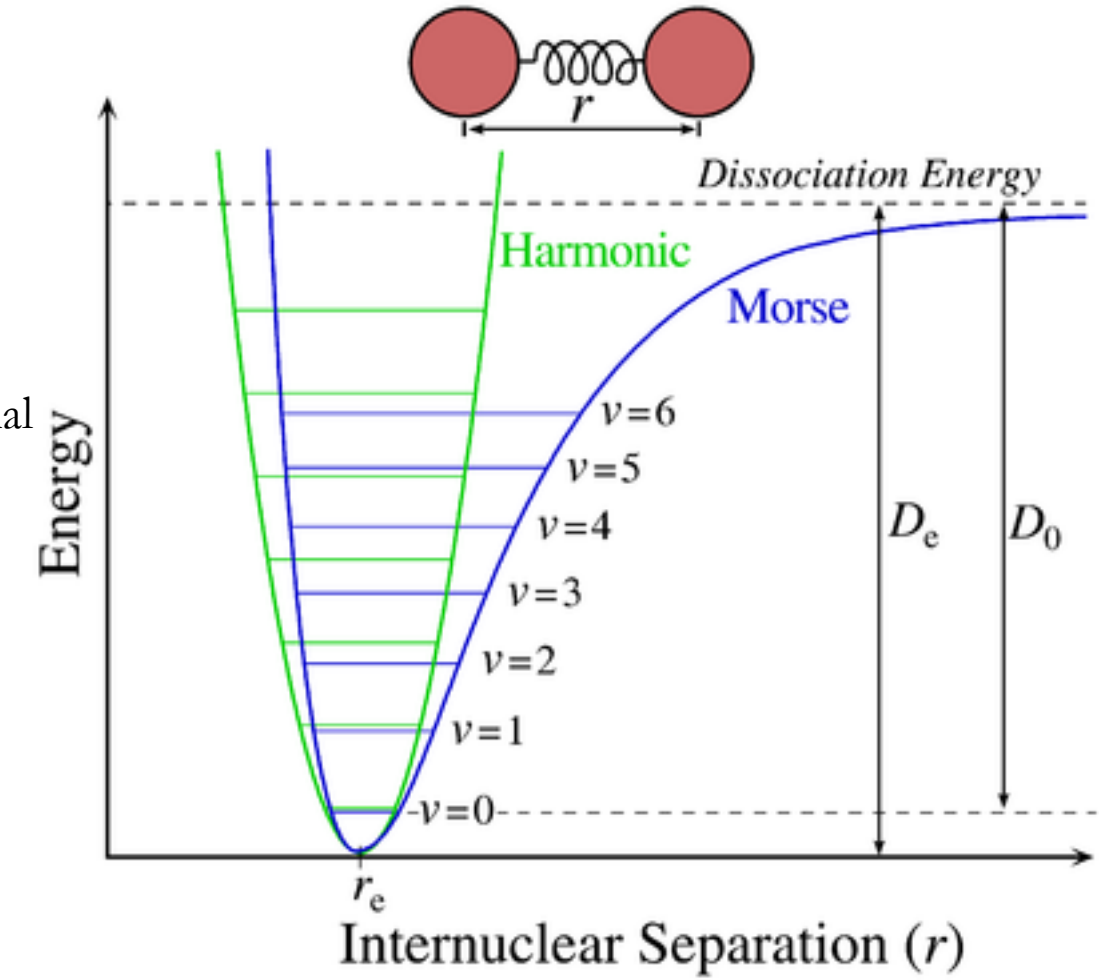
$$E = D_e(1-e^{-a(r-r_e)})^2$$

r: interatomic distance
r_e: equilibrium bond distance
D_e: well-depth
a: controls the width of the potential

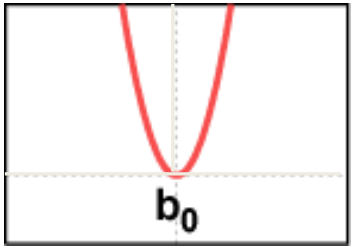
- Harmonic potential

$$E = \frac{1}{2}k(r-r_e)^2$$

r: interatomic distance
r_e: equilibrium bond distance
k: spring force constant

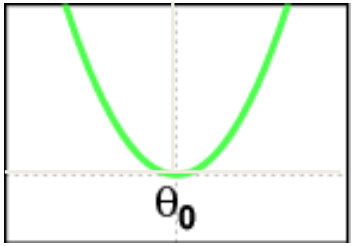


Bonded interactions



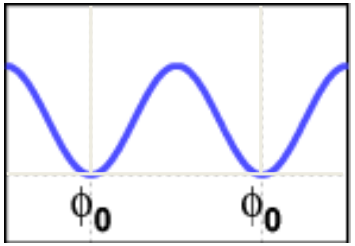
Bond

$$\sum_{bonds} K_r(\mathrm{r} - \mathrm{r}_{eq})^2$$



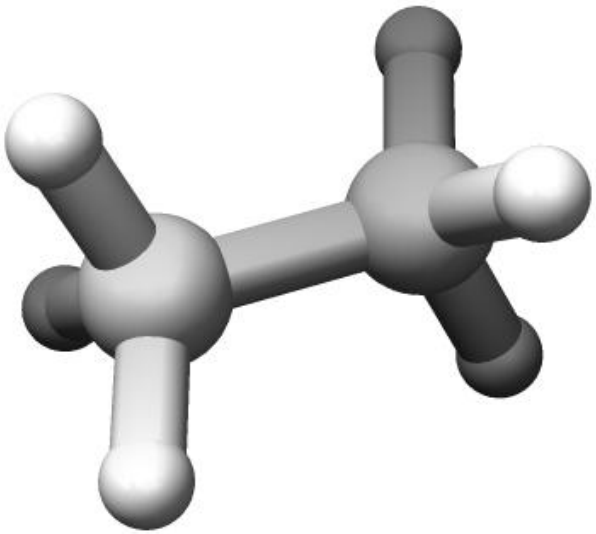
Angle

$$\sum_{angles} K_\theta(\theta - \theta_{\mathrm{eq}})^2$$

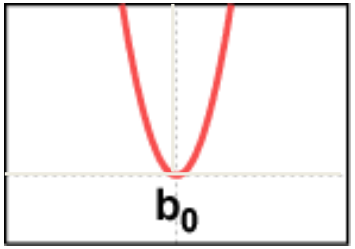


Dihedral

$$\sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)]$$

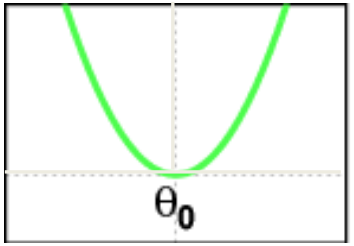


Bonded interactions



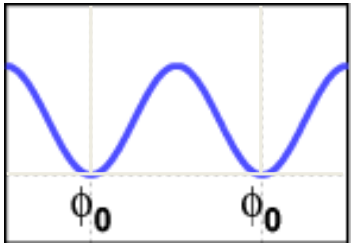
Bond

$$\sum_{bonds} K_r(\mathrm{r} - \mathrm{r}_{eq})^2$$



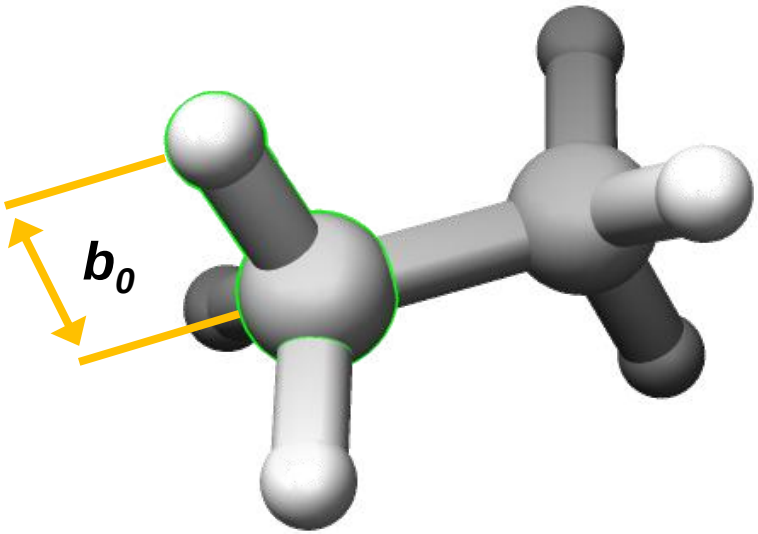
Angle

$$\sum_{angles} K_\theta(\theta - \theta_{\mathrm{eq}})^2$$

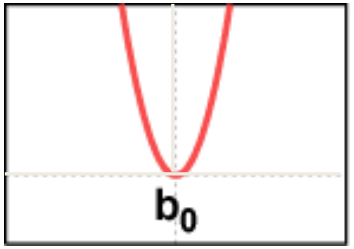


Dihedral

$$\sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)]$$

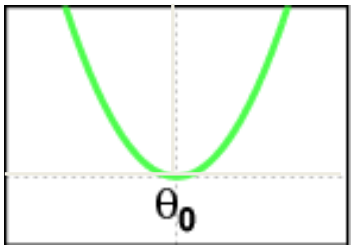


Bonded interactions



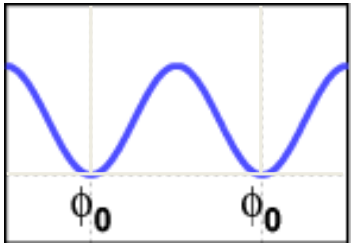
Bond

$$\sum_{bonds} K_r(\mathrm{r} - \mathrm{r}_{eq})^2$$



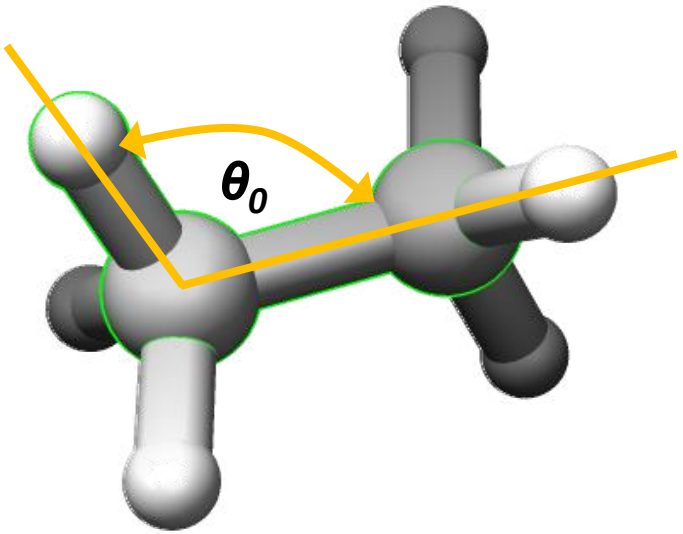
Angle

$$\sum_{angles} K_\theta(\theta - \theta_{\mathrm{eq}})^2$$

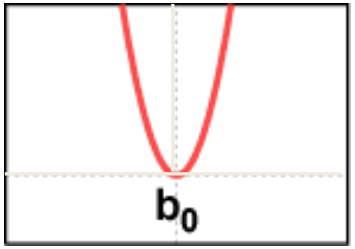


Dihedral

$$\sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)]$$

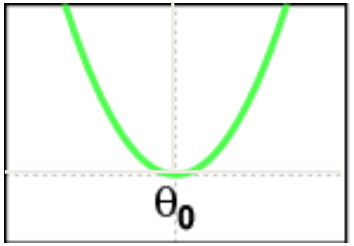


Bonded interactions



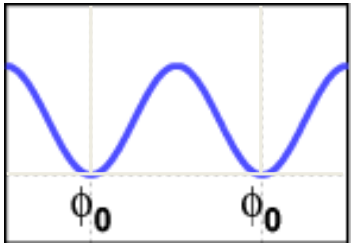
Bond

$$\sum_{bonds} K_r(\mathrm{r} - \mathrm{r}_{eq})^2$$



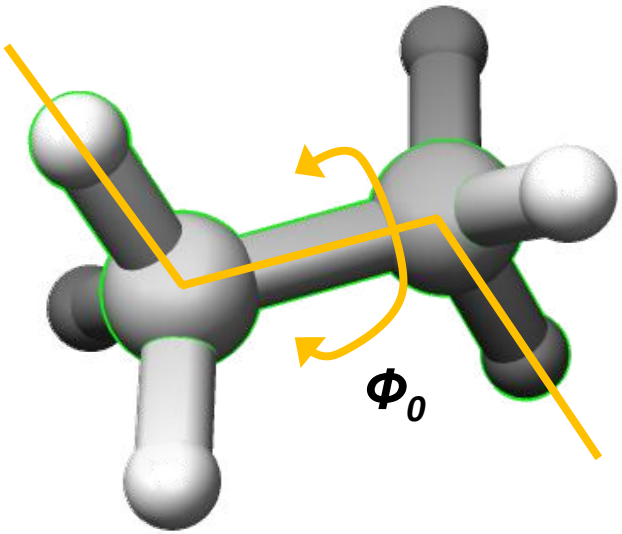
Angle

$$\sum_{angles} K_\theta(\theta - \theta_{\mathrm{eq}})^2$$

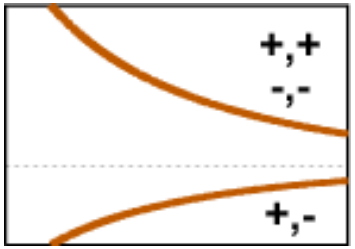


Dihedral

$$\sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)]$$

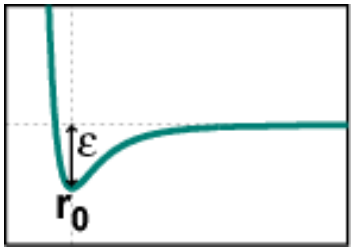


Non-bonded interactions



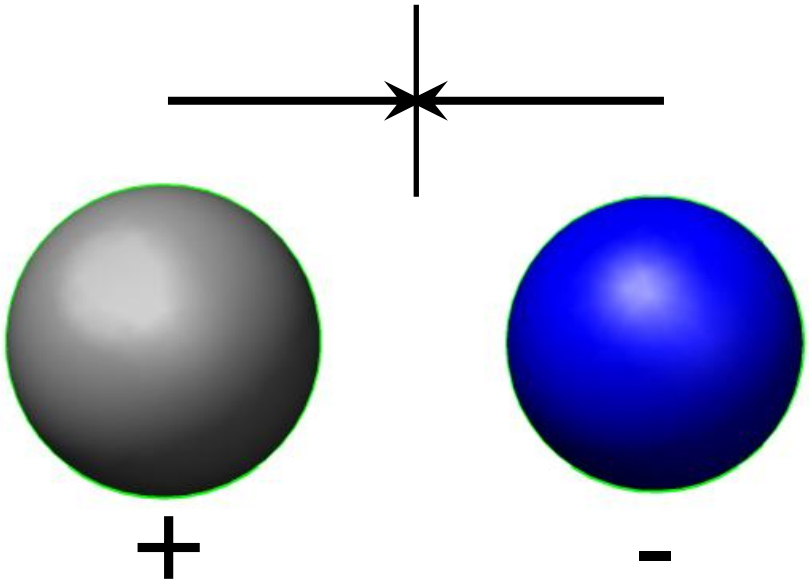
Electrostatics

$$\sum_{i<j}\left[\frac{q_i q_j}{\epsilon R_{ij}}\right]$$



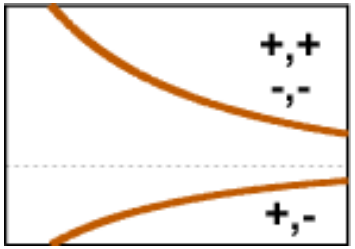
van der Waals

$$\sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}}-\frac{B_{ij}}{R_{ij}^{6}}\right]$$



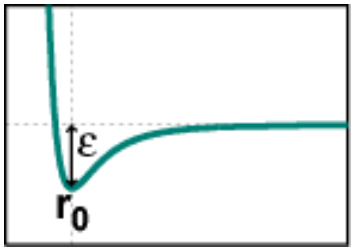
Coulomb interaction between single point charges

Non-bonded interactions



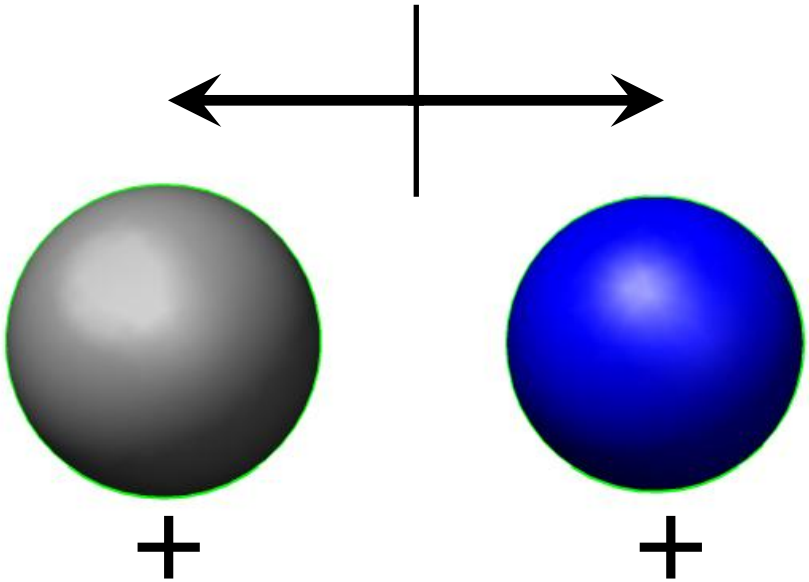
Electrostatics

$$\sum_{i<j}\left[\frac{q_i q_j}{\epsilon R_{ij}}\right]$$



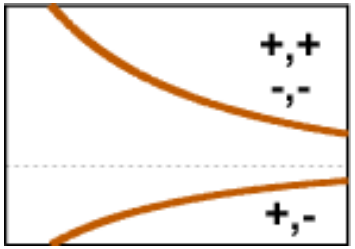
van der Waals

$$\sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}}-\frac{B_{ij}}{R_{ij}^{6}}\right]$$



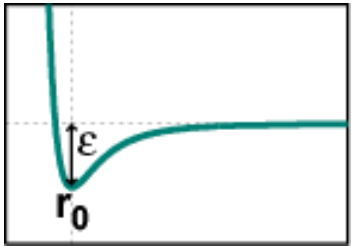
Coulomb interaction between single point charges

Non-bonded interactions



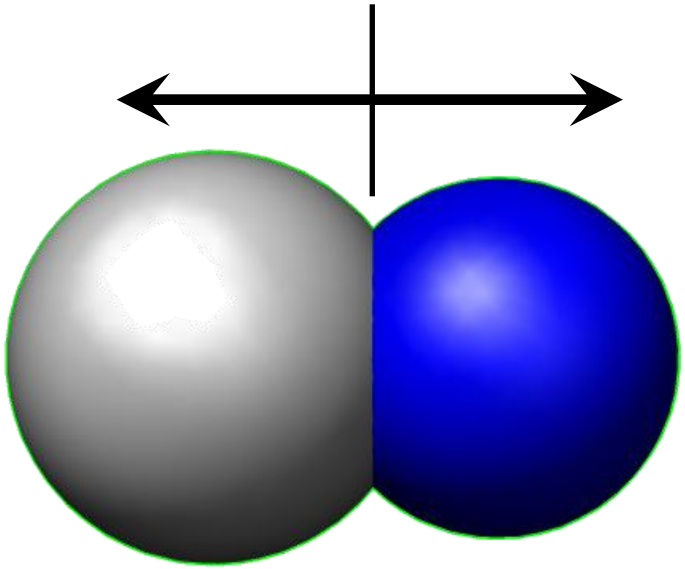
Electrostatics

$$\sum_{i<j}\left[\frac{q_i q_j}{\epsilon R_{ij}}\right]$$



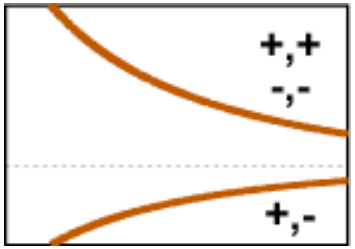
van der Waals

$$\sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}}-\frac{B_{ij}}{R_{ij}^{6}}\right]$$



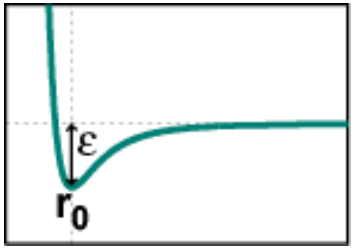
Hard core repulsion between close atoms

Non-bonded interactions



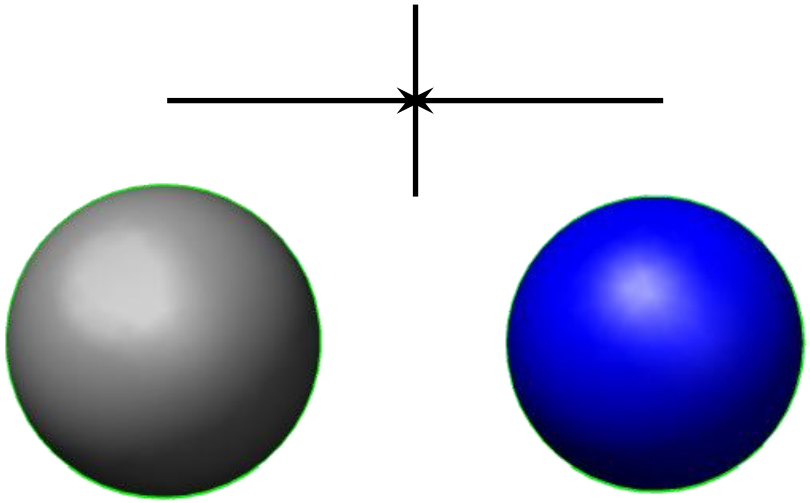
Electrostatics

$$\sum_{i<j}\left[\frac{q_i q_j}{\epsilon R_{ij}}\right]$$



van der Waals

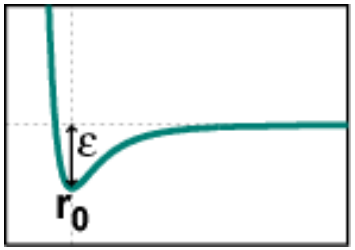
$$\sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}}-\frac{B_{ij}}{R_{ij}^{6}}\right]$$



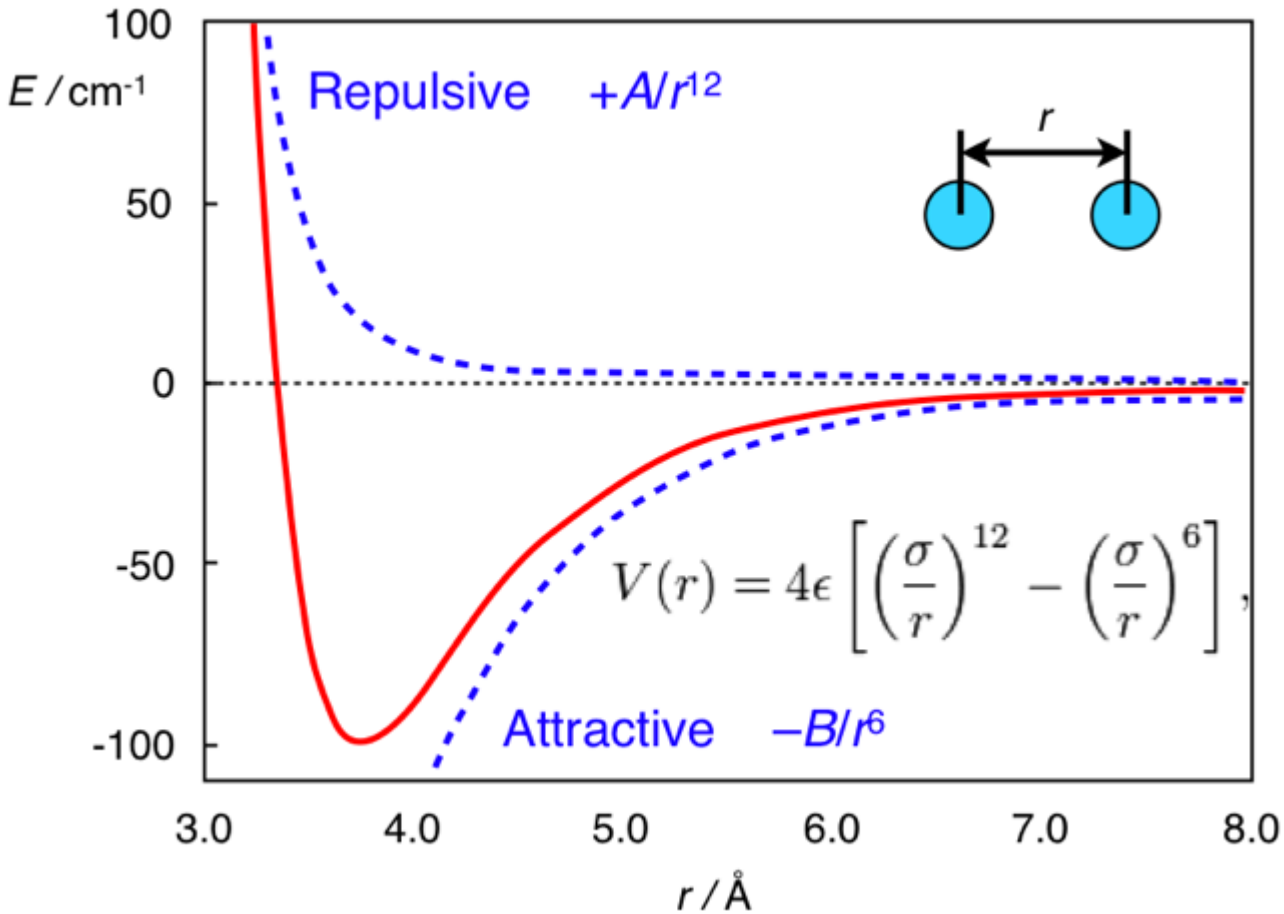
Weak dipole attraction between distant atoms

Lennard-Jones potential

van der Waals

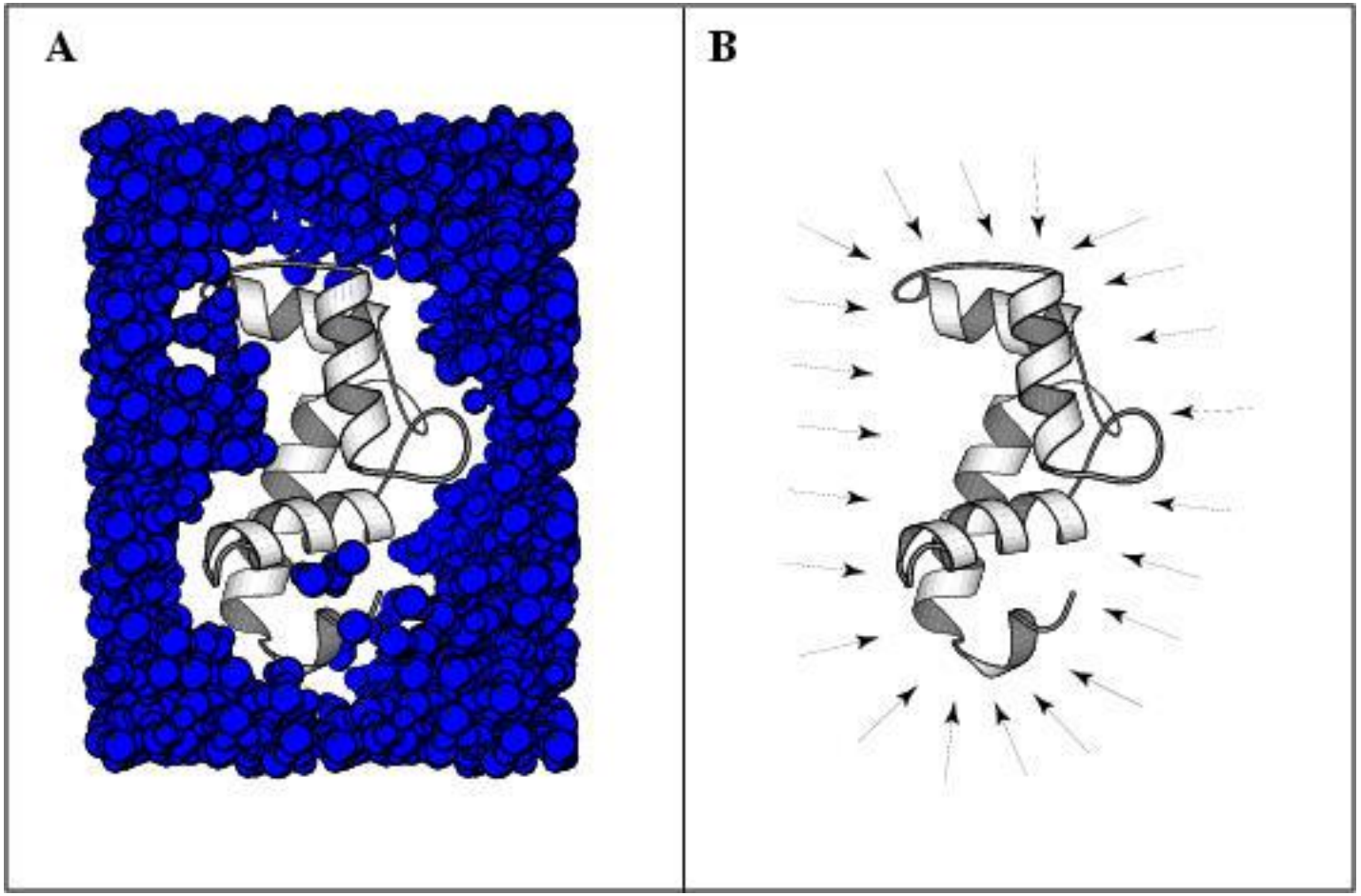


$$\sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}}-\frac{B_{ij}}{R_{ij}^{6}}\right]$$



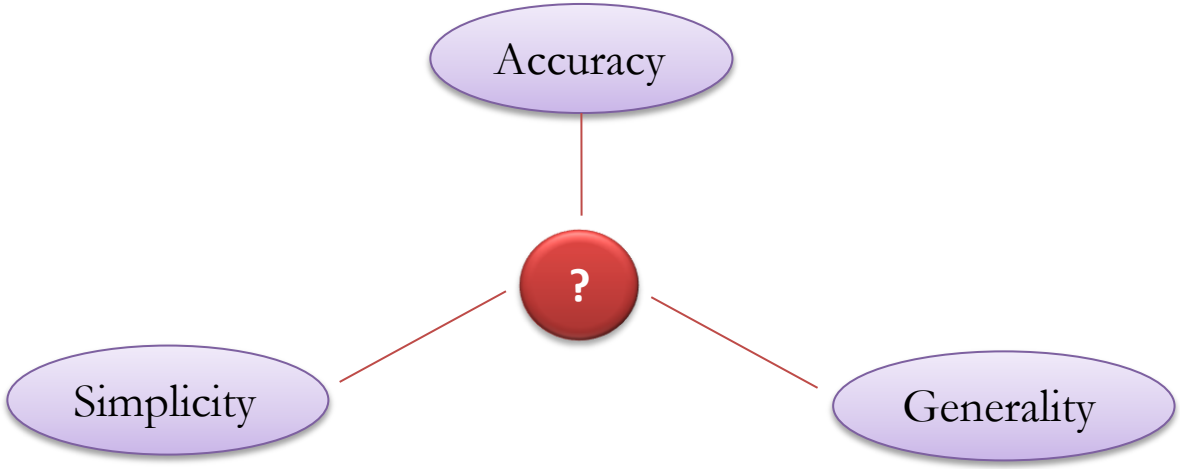
Solvent models

Explicit **Implicit**



Parametrization

Determining parameter values that best fit the force field and lead to the most accurate energy estimates is not trivial.



Parameters are fitted to experimental data (spectroscopy, small molecular crystals…) or quantum mechanics calculations. They are computed for a certain type of molecules (proteins, nucleic acids…) and may not be transferable.

Pair vs multi-body potentials

Coulombic and van der Waals potentials are summed over pairs of atoms.

How do we account for the influence of all the other particles in the system ?

Pairs: N(N-1)/2
Triplets: N(N-1)(N-2)/6

Effective potentials: account for the presence of the other entities through parametrization. An effective pair potential does not reflect the « true » interaction energy between two isolated atoms but is parametrized so as to include the effect of the other atoms in the energy of the pair.

Energy functions

Semi-empirical potentials

- analytical forms describing interactions which parameters are fitted to:
  - experimental data
  - quantum mechanics calculations

Statistical potentials

- analytical forms describing interactions which parameters are derived from a database of known structures
- sequence-structure association frequencies converted to free energies

Energy functions

Semi-empirical potentials

- analytical forms describing interactions which parameters are fitted to:
  - experimental data
  - quantum mechanics calculations

Statistical potentials

- **analytical forms describing interactions which parameters are derived from a database of known structures**
- **sequence-structure association frequencies converted to free energies**

Statistical mechanics

Proteins adopt an ensemble of **conformations** in solution. Not every protein in a large group of them has the lowest energy. The energies are random but they obey certain statistical laws based on the **Boltzmann distribution**.

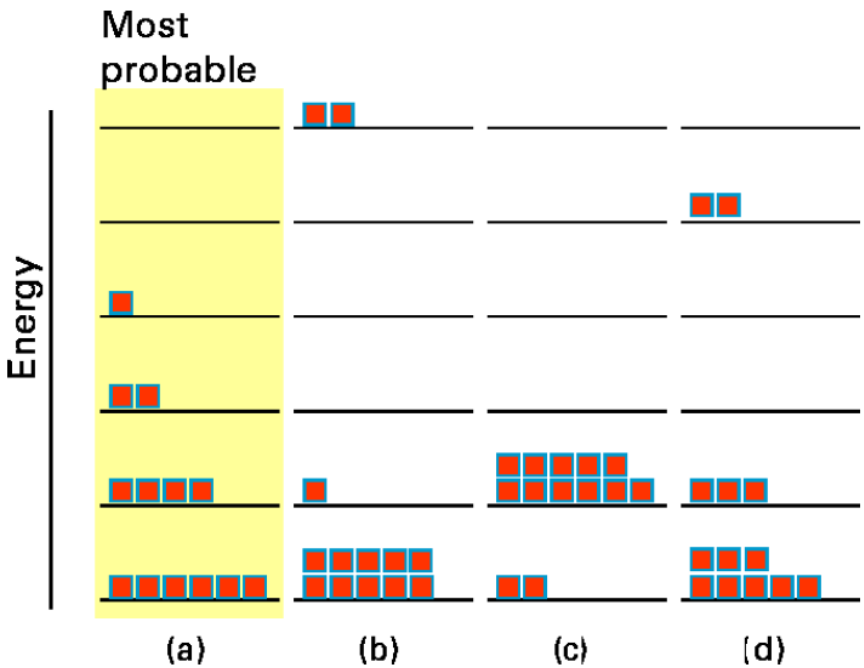
The probability of observing a given conformation C_i is:

Boltzmann coefficient

$$P(C_i) = \frac{\exp(-E(C_i)/kT)}{\sum \exp(-E(C_i)/kT)}$$

partition function

Boltzmann distribution

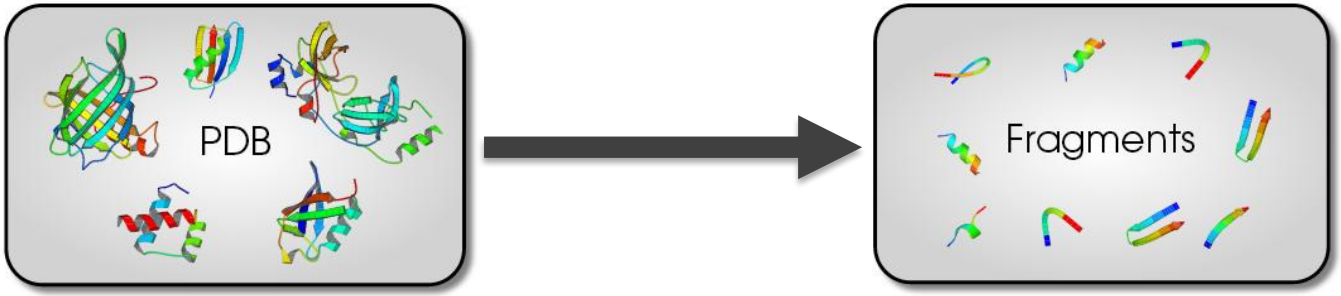


Statistical potentials: worflow

1/ Data Collection

- Low sequence similarity (<25 %) to avoid bias
- High resolution (<2 Å) and high quality structures
- Minimum critical size of the database to derive significant statistics

2/ Sequences and structures subdivision



1/ Parametrization of the potential

Distance , torsion, hydrophobicity…

Distance potentials

Potential of mean force $w^{(2)}$ between two particles at positions r_1 and r_2:

$$\exp[-w^{(2)}(\vec{r}_1,\vec{r}_2)/kT] = \frac{P^{(2)}(\vec{r}_1,\vec{r}_2)}{P^{(1)}(\vec{r}_1)P^{(1)}(\vec{r}_2)}$$

$P^{(1)}(r_1)$: probability of one particle being in position r_1
$P^{(2)}(r_1,r_2)$: probability of the two particles being in respective positions r_1 and r_2

Potential of mean force $W^{(2)}$ between two particles of types s_1 and s_2 at positions r_1 and r_2:

$$\exp[-W^{(2)}(\vec{r}_1,\vec{r}_2;s_1,s_2)/kT] = \frac{P^{(2)}(\vec{r}_1,\vec{r}_2\,|\,s_1,s_2)}{P^{(1)}(\vec{r}_1\,|\,s_1)P^{(1)}(\vec{r}_2\,|\,s_2)}$$

Distance potentials

Potential of mean force $\Delta W^{(2)}$ of a system with different types of particles compared to a reference system with only one type of particles:

$$\Delta W^{(2)}(\vec{r}_1,\vec{r}_2;s_1,s_2) = W^{(2)}(\vec{r}_1,\vec{r}_2;s_1,s_2) - w^{(2)}(\vec{r}_1,\vec{r}_2)$$

Free energy

denatured state

Estimation of the potential of mean force $\Delta W^{(n)}$ for the entire system by summing over all pairwise interactions:

$$\Delta W^{(n)}(\vec{r}_1,\ldots,\vec{r}_n;s_1,\ldots,s_n) = \sum_{i,j=1;i<j}^{n} W^{(2)}(\vec{r}_i,\vec{r}_j;s_i,s_j)$$

$s_1,\ldots,s_n$: amino acid types

$r_1,\ldots,r_n$: distance between amino acid residues

Distance potentials

Potential of mean force $\Delta W^{(2)}$ of a system with different types of particles compared to a reference system with only one type of particles:

In practice, how do we get the probabilities P and from there the free energy?

$$\Delta W^{(2)}(\vec{r}_{12}; s_1, s_2) = -kT \ln \frac{F(\vec{r}_{12} \mid s_1, s_2)}{F(\vec{r}_{12})}$$

Distances are computed between Cα, Cβ or side-chain centroïd Cμ or all atoms.

It is possible to introduced different levels of refinement by :

- combining frequencies of neighboring regions (potential smoothing)
- computing frequencies separately for aas close in the sequence (2-8 aas) and aas further than 8 aas (local/non-local potentials)

$s_1,\ldots,s_n$: amino acid types

$r_1,\ldots,r_n$: distance between amino acid residues

Energy functions

Semi-empirical potentials	
• Physical interpretation of the force field terms/parameters	• High cost to accurately account for solvent & entropic effects

Statistical potentials	
• Can be adapted to a coarse-grained representation of the protein • Implicitly include solvent & entropic effects	• No obvious physical interpretation • Dependence on some characteristics of the database

Energy functions evaluation

The performances of an energy function can be evaluated using **decoy sets**, generated by:

- Simulations of protein folding
- Comparative modeling
- Sequence inversion

Decoy sets must be **large**, contain **realistic** and **representative** structures.

A good energy function must:

- Assign the lowest energy to the native structure
- Discriminate the native structure from the decoys
- Display decrease of the energy of non-native structures as they become more and more **similar** to the native structures

→ Low RMSD, high coverage of native contacs

Conclusion

- **The native state of** a protein corresponds to the global free energy minimum. Proteins are only marginally stable.
- **The stability of a protein structure** can be expressed as a molecular mechanics potential energy or a potential of mean force (free energy)
- **The energy of stability** can be evaluated as a sum over physiscal terms describing the interatomic forces or over statistical terms describing the frequencies of co-occurrences of residue/atom conformations
- **Parameters** are either fitted to experimental data or more sophiscticate calculations, or derived from databases
- **Energy functions** must be carefully evaluated