Molecular Modeling and Simulation

Basic principles

Outlines

- Introduction – molecular dynamics

- Statistical mechanics

- Potential energy function

- Newton's law of motion and integrations algorithms

- Running MD simulations

- other simulation methods:

  - Langevin Dynamics

  - Brownian Dynamics

  - Monte Carlo

INTRODUCTION

WHAT IS MOLECULAR DYNAMICS?

- A computational method which describes equilibrium and dynamics properties of a biological system

- Generates configurations of the system by integration of Newton's laws of motion – calculate the time dependence of the molecular system

- Generates information at the microscopic level – atomic positions and velocities

- Connects structure and function by providing additional information to X-ray crystallography and NMR

Historical Background

- Introduced by **Alder and Wainwright** in 1950's to study interaction of hard spheres (atoms that interact through perfect collisions)
- First simulation carried out by **Rahman** (1964) for liquid Argon
- First protein simulation carried out by **McCammon** (1977) for bovine pancreatic tyrpsin inhibitor (BPTI)
- Today – solvated proteins, protein – DNA complexes, lipid systems, etc.

Statistical Mechanics

- In molecular dynamics simulation we explore the **macroscopic** properties of a system through **microscopic** simulations

- The connection between microscopic simulations and macroscopic properties is made via **statistical mechanics**, which studies a macroscopic system from a molecular point of view

- The distribution of the system within the ensemble follows **Bolzmann** distribution

Definitions

- **Ensemble** – collection of all possible systems which have different microscopic states but identical macroscopic or thermodynamic state

- **Microcanonical ensemble** (**NVE**): the thermodynamic state is characterized by fixed number of atoms N, fixed volume V and fixed energy E.

- **Canonical ensemble** (**NVT**)

- **Grand canonical** (μVT) where μ is chemical potential

Calculating averages from MD simulations

- Need to consider the average since the sample contains a large number of conformations
- Average values in statistical mechanics corresponds to ensemble averages
- Calculating this requires integrating over all possible states of the system which is extremely difficult
- **Ergodic hypothesis** states that the time averages equal the ensemble average

$$\langle A \rangle_{ensemble} = \langle A \rangle_{time}$$

The basic idea is to allow the system to evolve in time so that the system will eventually pass through all possible states
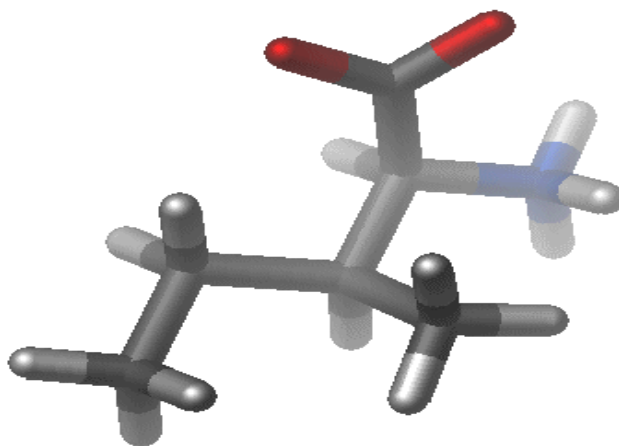
Potential Energy function

Force Fields

The effective molecular energy can be expressed as a sum of potentials derived from simple physical forces

$$V = E_{bonded} + E_{nonbonded}$$

$$E_{bonded} = E_{bond\text{-}strech} + E_{angle\text{-}bend} + E_{rotate\text{-}along\text{-}bond}$$

$$E_{nonbonded} = E_{van\text{-}der\text{-}Waals} + E_{electrostatic}$$

Bond length potentials

$$E_{bond-stretch} = \sum_{1,2\,pairs} K_b \left(b - b_0 \right)^2$$

- Molecules undergo vibrational motion, which is modeled as harmonic potential according to **HOOKE's** law
- Kb represents the force constant:
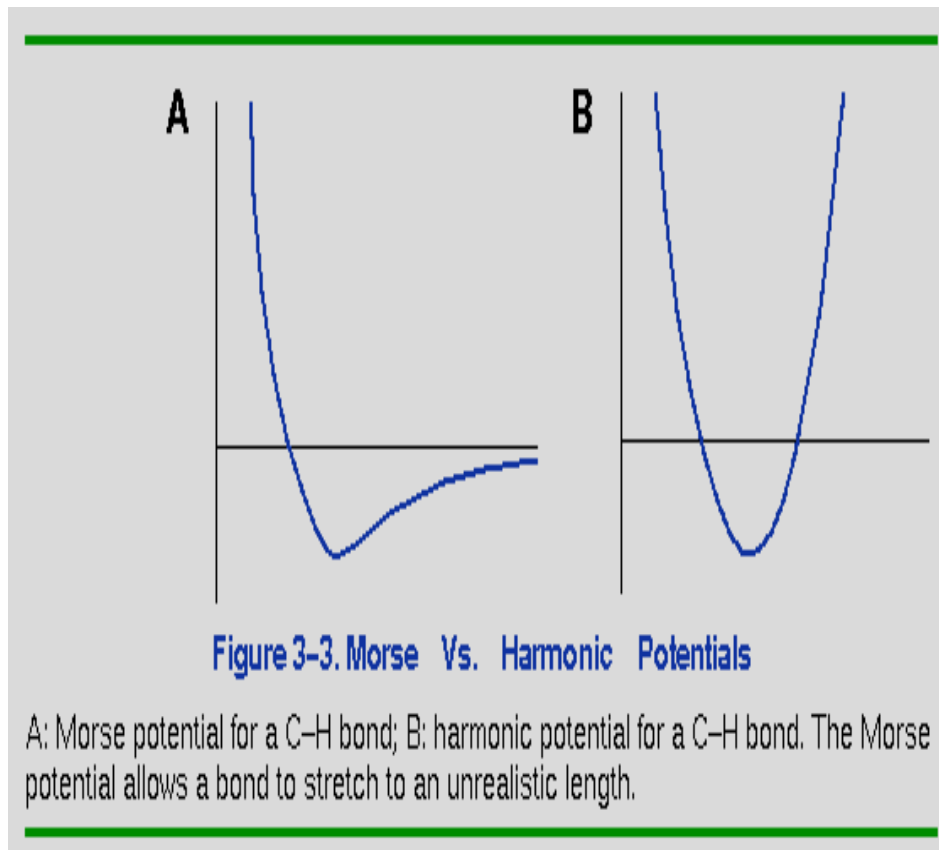
$$K_b = \mu \omega^2$$

$$\mu = \frac{m_1 m_2}{m_1 + m_2}$$

- This harmonic potential is valid only for deviations of 0.1Å or less. At larger distances the atoms dissociate and no longer interact.

Harmonic term vs. Morse term

Morse term is more accurate, however it is generally not used in MD simulations since it requires 3 parameters to be specified for each bond

$$v(l) = D_e \left\{ 1 - \exp\left[-a(l - l_0) \right] \right\}^2$$

Figure 3–3. Morse Vs. Harmonic Potentials

A: Morse potential for a C–H bond; B: harmonic potential for a C–H bond. The Morse potential allows a bond to stretch to an unrealistic length.

Bond angle potentials

$$E_{bond-bend} = \sum_{angles} K_{\theta}(\theta - \theta_0)^2$$

- Describe the deviation from an ideal bond angle geometry
- K_{θ} represent angle bending constant, θ_0 represent the deviation from the ideal bond angle

Torsion angle potential

$$E_{rotate-along-bond} = \sum_{1,4\,pairs} K_\phi \left(1 - \cos(n\phi)\right)$$

○ This terms models the steric barrier between atoms separated by 3 covalent bonds

○ The motion associated is rotation, described by a dihedral angle around the middle bond

○ The potential is assumed to be periodic and expressed as a cosine function

○ K_Φ represent rotation constant, n represent the periodicity of the rotational barrier and Φ the dihedral angle

The Coulomb potential

$$V_{elec} = \frac{q_i q_j}{\varepsilon_{eff} \, r_{ij}}$$

- Coulombic interaction decay slowly with distance, considered long range interactions

- **r** represents the distance between two atoms having charges **q_i** and **q_j**

- **ε** represent the dielectric constant, a number relating the ability of a material to carry alternating current to the ability of vacuum to carry alternating current

The Van-der-Waals term

$$E_{van-der-Waals} = \sum_{\substack{nonbonded \\ pairs}} \left(\frac{A_{ik}}{r_{ik}^{12}} - \frac{C_{ik}}{r_{ik}^{6}} \right)$$

Best described as a Lenard-Jones potential
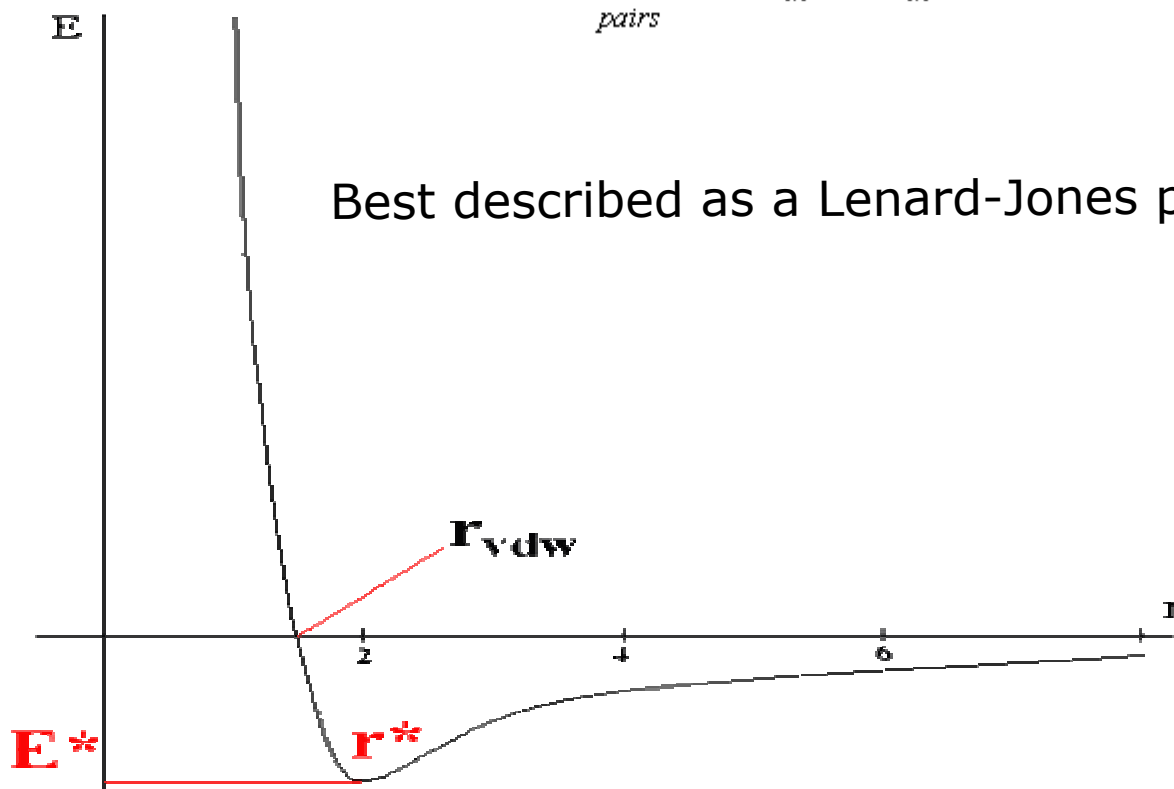
The Van-der-Waals term

- Express the interaction energy between two atoms
- Contains an attractive part and a repulsive part
- Attractive forces due to London forces (dipole – dipole interaction)
- Repulsive part due to Pauli-exclusion principle and inter-nuclear repulsion
- A,C parameters can be determined by a variety of methods

Treatment of the non-bonded energy terms

- The calculation of the non-bonded part is the most time consuming

- To speed up the computations use a cutoff distance, which is pre defined such that a distances greater than it are ignored
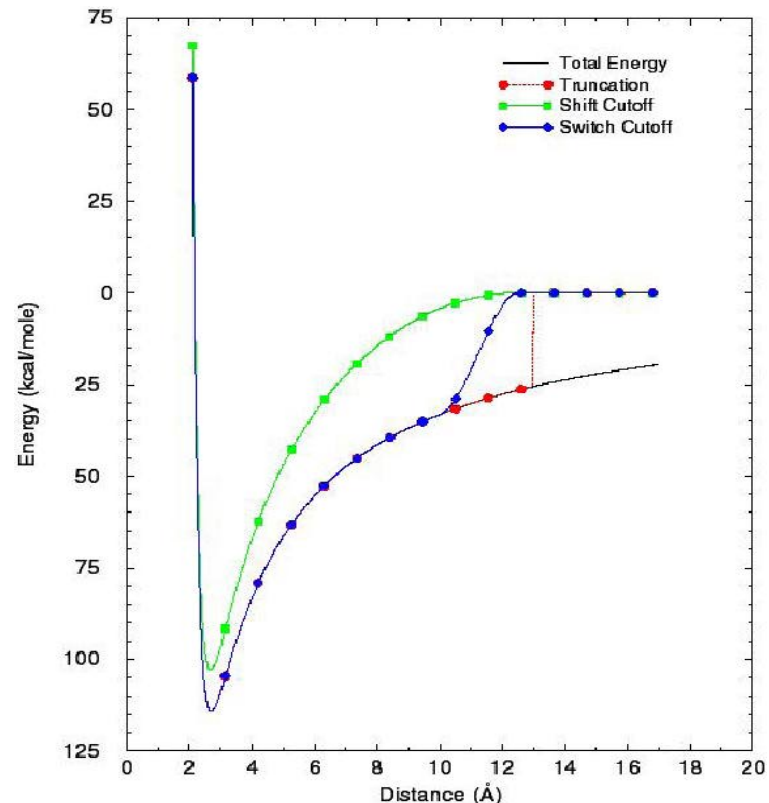
Newton's laws of motion

Integration algorithms

Classical mechanics

- The molecular dynamics simulation is based on Newton's law of motion: **F=ma**

- By knowing the force that is acting on each atom it is possible to determine the acceleration of each atom. Integration of the equation yields a trajectory that describes the positions, velocities and acceleration of each atom as they are varied with time

- Once the positions and velocities of each atom are known, the state of the system can be predicted at any time

Some mathematics…

- Newton's equation of motion

$$F = ma$$

- The force can be written as the gradient of the potential energy

$$F = -\nabla V$$

- Combine the two equations to get

$$\frac{dV}{dr} = -m\frac{d^2 r}{dt^2}$$

A trajectory is obtained by solving this differential equation

Basic idea

1. Integration broken down to many small stages: δt

2. The total force on each particle in the configuration at a time t is the vector sum of its interactions with other particles.

3. From the force determine the acceleration of the particles and combine it with positions and velocities at time t to calculate at time $t + \delta t$

The force is constant during the time step!

How to integrate equation of motion?

○ The potential energy is a function of the atomic positions of all the atoms in the system.

○ Due to this complexity there is no analytical solution

⟹ Use algorithms that approximate the positions, velocities and accelerations with **Taylor's expansion**

$$x(t+\Delta t)=x(t)+\frac{dx(t)}{dt}\Delta t+\frac{d^2x(t)}{dt^2}\frac{\Delta t^2}{2}+...$$

Examples of algorithms techniques

Common use is the **VERLET** algorithm:

- Each position is determined from the current position and position at time $t - \delta t$

$$x(t + \delta t) = 2x(t) - x(t - \delta t) + a(t)\delta t^2$$

- Advantage – straightforward, modest storage requirements
- Disadvantages – not accurate, lack of explicit velocity term

Examples of algorithm techniques - continue

Leap-frog algorithm

- Calculates positions at time t and velocities at time $t - \dfrac{\delta t}{2}$

$$x(t + \delta t) = x(t) + v(t + \dfrac{\delta t}{2})\delta t$$

$$v(t + \dfrac{\delta t}{2}) = v(t - \dfrac{\delta t}{2}) + a(t)\delta t$$

- Advantage – the velocities are calculated explicitly
- Disadvantage – velocities not calculated at the same time as positions, therefore cannot calculate kinetic energy contribution to the total energy at the same time as the positions are defined

Problems

The most demanding part of MD simulation is calculating the force acting on each atom, need to find methods to limit the number of force calculations:

- Select an integrator to allow longer time steps. however, δt Should be small compared to the period of the highest frequency!

- **SHAKE** algorithm – impose constraints on the bond lengths. It averages out the highest frequency vibrations

- Use reduced representation of the system, for example **GO** model

- Use reduced number of interactions, for example implicit solvent and periodic boundaries.

Time scales of motions

- **Local Motions** (0.01 to 5 Å, 10^{-15} to 10^{-1} s)
  - Atomic fluctuations
  - Side-chain Motions
  - Loop Motions
- **Rigid Body Motions** (1 to 10Å, 10^{-9} to 1s)
  - Helix Motions
  - Domain Motions (hinge bending)
  - Subunit motions
- **Large-Scale Motions** (> 5Å, 10^{-7} to 10^{4} s)
  - Helix coil transitions
  - Dissociation/Association
  - Folding and Unfolding

Running MD simulation

Experimental conditions

- Early MD simulations were performed in the **NVE** ensemble, recent methods perform also at **NTP**

- during a simulation at constant energy the temperature will fluctuate. The temperature can be calculated from the velocities using the relationship:

$$E_k = \frac{3}{2} N K_B T$$

$$T = \frac{1}{3 N K_B} \sum_{i=1}^{N} m_i v_i^2$$

Constant temperature

○ Couple the system to an external heat bath that is fixed at the desired temperature

○ The bath acts as a source of thermal energy, supplying or removing heat from the system as needed.

○ The velocities are scaled at each temp. such that:

$$\frac{dT(t)}{dt} = \frac{1}{\tau}(T_{bath} - T(t))$$

Constant temperature

τ is a coupling parameter which determines how tightly the bath and the system are coupled together. Large τ means weak coupling and vice versa.

In order to alter the temperature the velocities need to be scaled, the scaling factor is

$$\lambda^2 = 1 + \frac{\delta t}{\tau}(\frac{T_{bath}}{T(t)} - 1)$$

Steps taken to setup and run MD simulation:

A molecular dynamics simulation consists of 3 essential parts:

- **Initialization**
- **Equilibrium**
- **Production** - Calculate thermodynamic properties of the system

Initialization

- Specify the initial coordinates and velocities
- Initial structure is obtained by x-ray crystallography or NMR structure
- Velocities are chosen randomly from Maxwll-Bolzmann or Gaussian distribution:

$$\rho(v_{ix}) = \left(\frac{m_i}{2\pi K_b T} \right)^{\frac{1}{2}} \exp\left(-\frac{1}{2} \frac{m_i v^2_{ix}}{K_b T} \right)$$

Energy minimization

- Prior to starting a simulation it is advisable to do energy minimization

- Useful in correcting flaws such as steric clashes between atoms and distorted bond angles/lengths

- Need to find the minimum value for the energy function with 3N degrees of freedom

Energy minimization

It is a nonlinear task with many local minima

There are several techniques, for example

- **Steepest descent** – relates to the gradient of the potential energy surface:

$$F(r) = -\frac{dV(r)}{dr}$$

  the force acts as a guide toward the local energy minima.

  It is an iterative approach – after every iteration the step size is adjusted according to whether the potential energy had decreased/increased.

  inefficient for problems with multiple local minima.

Energy minimization

Another technique is the **Newton-Raphson**:

It assumes that in the region of the minima the potential energy is quadratic:

$$V(x_i) \approx a + bx_i + cx_i^2$$

the minima is calculated using the derivative

$$\frac{dV(x_{min})}{dx} = 0$$

Usually not very useful since most potentials are not quadratic.

Solvation

- The solvent (usually **water**) has a fundamental influence on the structure, dynamics and thermodynamics of biological molecules
- One of the most important effects of the solvent is the screening of electrostatic interactions
- There are several methods for the treatment of the solvent effect:
  - **implicit** treatment of the solvent: an effective dielectric constant is used. Often the effective dielectric constant is taken to be distance dependent
  - Periodic boundary conditions

Periodic Boundary conditions

- Periodic boundary conditions enable a simulation to be performed using a relatively small number of particles in such a way that the particles experience forces as though they were in a bulk solution
- The box containing the system is replicated infinitely in all directions to give a periodic array
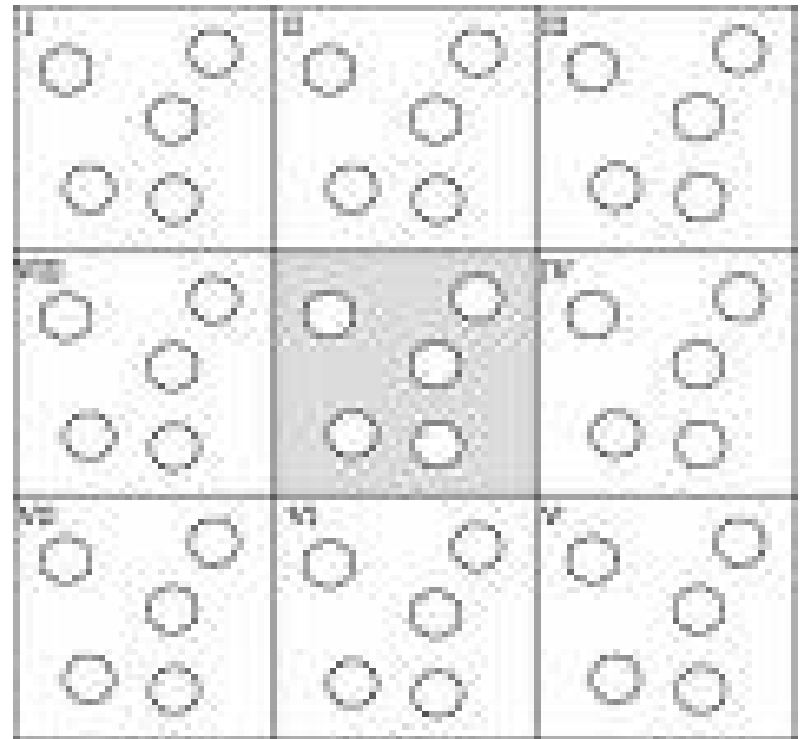
Equilibrium phase

- During this phase several properties are monitored: pressure, temperature and energy

- The purpose of this phase is to run the simulation until these properties become stable with time

Simulation methods

Langevin Dynamics

- The Langevin equation is a ***stochastic*** differential equation in which two force terms have been added to Newton's second law

- A molecule in the real world is not present in a vacuum. Jostling of solvent molecules causes friction and there are also collisions that perturb the system

- The model was employed to avoid explicit representation of water molecules, enhance sampling, represent hydration shell models in large system and other uses.

Langevin equation

$$m_i \frac{d^2 r}{dt^2} = F_i(r) - M\gamma \frac{dr}{dt} + R_i(t)$$

- ○ The effects of solvent molecules can be approximated in terms of a frictional drag on the solute as well as random kicks associated with the thermal motions of the solvent molecules
- ○ The molecules interact with a stochastic heat bath via random forces and dissipative forces

Langevin parameters

γ represent the **collision frequency**: controls the magnitude of the frictional force and the variance of the random forces.

it ensures that the system converges to Bolzmann distribution.

The larger its value, the greater the influence of the surrounding fluctuating force (solvent). Small γ implies inertial motion.

If the main objective is to control temperature, one need to use small values of γ

Langevin parameters

3 different situations can be considered, depending on the magnitudes of the integration time step and γ :

1. $\gamma \delta t \leq 1$: small influence of the solvent
2. $\gamma \delta t \geq 1$: diffusion limit
3. Intermediate

for each of those limits can find algorithms to solve the differential equations.

Langevin parameters

R(t) is a random force vector, which is a stationary Gaussian process with zero-mean: $\langle R(t) \rangle = 0$

$$\langle R(t)R(t^{'})^{T} \rangle = 2\gamma K_b TM\delta(t - t^{'})$$

Dirac δ function:

$$\delta = \begin{cases} \infty, t = 0^{'} \\ 0, t \neq 0^{'} \end{cases}$$

$$\int_{-\infty}^{\infty} f(t)\delta(t)dt = 1$$

The random force is chosen independently each step

Advantages of langevin Dynamics

- The number of degrees of freedom is the number of solute which leads to a reduction in time of calculations. This allows to perform **longer simulations**

- The random motions can assist in propagating **barrier-crossing motions** and therefore can improve conformational sampling characteristics

- Efficient in simulations of **long chain molecules** such as DNA with many base pairs

Brownian dynamics

- Describes the **diffusive** limit of the langevin equation where the motion is more **random**
- Assumes a high viscosity of the fluid such that the solute reorients continuously by colliding with the solvent molecules
- Uses a very **large friction coefficient**
- Useful in describing processes such as diffusion controlled reactions

Monte Carlo

○ Generates configurations of the system by making **random changes**

○ **Algorithm**: calculate the energy of the new configuration

- if $\triangle$**E<0** then the configuration is accepted
- if $\triangle$**E>0** then it will depend on Bolzmann factor.

This process is iterated until sufficient sampling statistics for the current temperature are achieved.

Comparison between simulation methods

method	applications and advantages	disadvantages
Molecular Dynamics	Continuous motion, experimental bridge between structures and macroscopic kinetic data	Short time span
Langevin Dynamics	Long-time trajectories	High CPU
Brownian Dynamics	Large scale and long-time motion	Limited to systems with small inertia
Monte Carlo	Large scale sampling, useful statistics	Difficult to devise structural perturbations
Energy minimization	Correcting structural flaws	No dynamics information