Biomolecular modeling I

Marcus Elstner and Tomáš Kubař

2014, December 16

Molecular mechanics

classical description of molecules

Motivation

- To investigate the function of biomolecules,
  we need to characterize its structure and dynamics.
- We will look how the molecules are moving
  – Molecular Dynamics

Motivation

- To investigate the function of biomolecules,
  we need to characterize its structure and dynamics.
- We will look how the molecules are moving
  – Molecular Dynamics
- For this, we need to calculate the energy of the system
  and the forces on atoms

Motivation

- To investigate the function of biomolecules,
  we need to characterize its structure and dynamics.
- We will look how the molecules are moving
  – Molecular Dynamics
- For this, we need to calculate the energy of the system
  and the forces on atoms
- Energy from quantum mechanics / quantum chemistry
  seemingly easy $E = \left\langle \Psi \middle| \hat{H} \middle| \Psi \right\rangle$

Motivation

- To investigate the function of biomolecules,
  we need to characterize its structure and dynamics.
- We will look how the molecules are moving
  – Molecular Dynamics
- For this, we need to calculate the energy of the system
  and the forces on atoms
- Energy from quantum mechanics / quantum chemistry
  seemingly easy $E = \left\langle \Psi \middle| \hat{H} \middle| \Psi \right\rangle$
  but not quite possible for large molecular systems

Motivation

$$\left[-\frac{\hbar^2}{2m_e\lambda^2}\nabla'^2 - \frac{e^2}{4\pi\varepsilon_0\lambda r'}\right]\phi' = \mathscr{E}\phi'$$

$$[ij|kl] = \int d\mathbf{x}_1\, d\mathbf{x}_2\, \chi_i^*(\mathbf{x}_1)\chi_j(\mathbf{x}_1)r_{12}^{-1}\chi_k^*(\mathbf{x}_2)\chi_l(\mathbf{x}_2)$$

$$\begin{aligned} F_{\mu\nu} &= \int d\mathbf{r}_1\, \phi_\mu^*(1)f(1)\phi_\nu(1) \\ &= \int d\mathbf{r}_1\, \phi_\mu^*(1)h(1)\phi_\nu(1) + \sum_a^{N/2} \int d\mathbf{r}_1\, \phi_\mu^*(1)[2J_a(1) - K_a(1)]\phi_\nu(1) \\ &= H_{\mu\nu}^{\mathrm{core}} + \sum_a^{N/2} 2(\mu\nu|aa) - (\mu a|a\nu) \end{aligned}$$

$$|\Phi\rangle = c_0|\Psi_0\rangle + \sum_{ra} c_a^r|\Psi_a^r\rangle + \sum_{\substack{a<b\\ r<s}} c_{ab}^{rs}|\Psi_{ab}^{rs}\rangle + \sum_{\substack{a<b<c\\ r<s<t}} c_{abc}^{rst}|\Psi_{abc}^{rst}\rangle + \cdots$$

$$E_0^{(2)} = \sum_{a,b=1}^{N/2} \sum_{r,s=(N/2+1)}^{K} \frac{\langle ab|rs\rangle(2\langle rs|ab\rangle - \langle rs|ba\rangle)}{\varepsilon_a + \varepsilon_b - \varepsilon_r - \varepsilon_s}$$

Idea of molecular mechanics

- often – well localized bonding orbitals (organic molecules)
- idea – similar bonds have similar strength and properties
  e.g. similar C–H σ-orbitals → all C–H bonds are 'similar'

- possibly by a harmonic spring? (the simplest possible function)

$$\begin{aligned} E(x) &= \frac{1}{2}k(x-x_0)^2 \\ F(x) &= -\frac{\partial E(x)}{\partial x} = -k(x-x_0) \end{aligned}$$

- 2 parameters k and x_0 with defined meaning
  – can be obtained from spectroscopy

Concept of (atom, bond. . .) type

- let us use harmonic springs for covalent bonds
- we do not want to parametrize k and x_0 each bond separately
- use just several sets (k, x_0), for different types of bonds

Why can we expect such ‘unification’ to work?

Concept of (atom, bond. . .) type

Spectroscopy

- every C–H bond: length 1.06–1.11 Å,
  frequency ca. 3100 cm^{-1}, in any molecular environment

Thermochemistry

- heat of formation – roughly additive:
  $CH_4 \cong 4$ C–H
  $C_2H_6 \cong 6$ C–H + C–C

Concept of (atom, bond. . .) type

How to identify the atom types? – chemical ideas

Concept of (atom, bond. . .) type

How to identify the atom types? – chemical ideas

i) hybridization

- different types for sp^3 carbon (4 bonds) and sp^2 C (3 bonds)
- different functions for bonds of types C–C, C=C and C≡C
- determine the parameters (k, x_0) with some selected molecules, typical for the binding situation
- example: use C_2H_6, C_2H_4, C_2H_2 and benzene for k, x_0

Concept of (atom, bond. . .) type

How to identify the atom types? – chemical ideas

ii) polarity

- an atom bonded to electronegative atom – electron deficient
  → affects its bonding to other atoms
- example: C–C bond in O=CH–C. . . is affected
  and needs to be parametrized differently from apolar C–C
  → an atom type for carbonyl C introduced

Biomolecular force fields
– usually 20 types for C, 10 for N and 5 for O and H

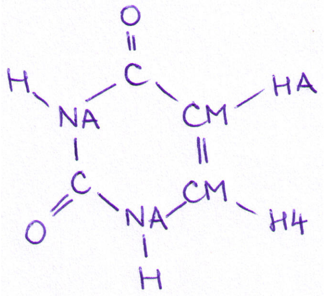
Concept of (atom, bond. . .) type

AMBER types for carbon:

C - sp2 C carbonyl group
CA - sp2 C pure aromatic (benzene)
CB - sp2 aromatic C, 5&6 membered ring junction
CC - sp2 aromatic C, 5 memb. ring HIS
CK - sp2 C 5 memb.ring in purines
CM - sp2 C pyrimidines in pos. 5 & 6
CN - sp2 C aromatic 5&6 memb.ring junct.(TRP)
CQ - sp2 C in 5 mem.ring of purines between 2 N
CR - sp2 arom as CQ but in HIS
CT - sp3 aliphatic C
CV - sp2 arom. 5 memb.ring w/1 N and 1 H (HIS)
CW - sp2 arom. 5 memb.ring w/1 N-H and 1 H (HIS)
C* - sp2 arom. 5 memb.ring w/1 subst. (TRP)

Concept of (atom, bond. . .) type

AMBER atom types in a molecule of uracil



Interactions between atoms

Bonded

- mediated by, and resulting directly from covalent bonds
- cover all of the quantum-mechanical phenomena between pairs of atoms with effective potentials
- harmonic springs between atoms (also angles and dihedrals)

Interactions between atoms

Bonded

- mediated by, and resulting directly from covalent bonds
- cover all of the quantum-mechanical phenomena between pairs of atoms with effective potentials
- harmonic springs between atoms (also angles and dihedrals)

Non-bonded

- longer-range interactions
  – charge–charge (Coulomb) and van der Waals (vdW)
- between molecules and distant parts of one molecule

Coulomb interaction

- idea – condense electrons in each atom with the nucleus
  → effective atomic charge $q_i = -Q_i + Z_i$:

$$E_{QQ} = \frac{1}{2} \sum_{ij} \frac{q_i \cdot q_j}{R_{ij}}$$

- needs to be defined for every atom
  rather than atom type – this would be too crude
- from quantum-chemical calculations
  of typical (bio)molecular fragments
  – amino acid residues and peptide bonds for proteins
  – nucleobases, sugars and phosphate groups for DNA/RNA

Coulomb interaction

How to calculate atomic charges?

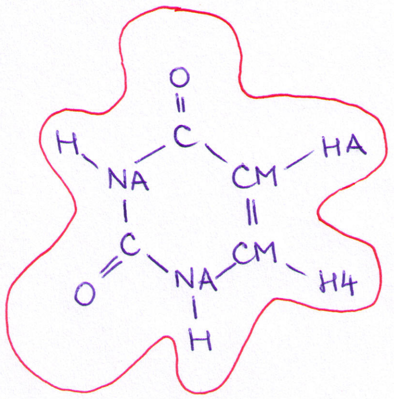
- popular – potential-derived charges:
    1. calculate the electron density in the molecule
    2. get electrostatic potential at surface of the molecule
    3. fit point electric charges on atoms to reproduce the ESP

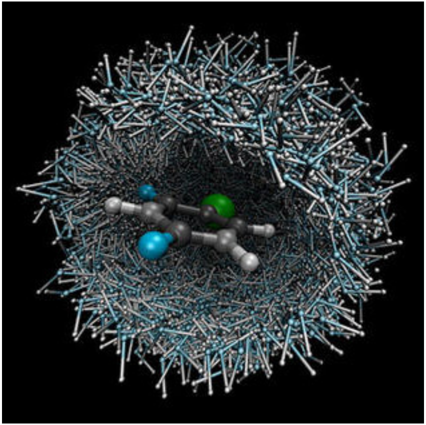
Possible improvement – polarizable force field

- atomic polarizability α_i is assigned to every atom *i*
- external field induces atomic dipole $\overrightarrow{\mu_i} = \overleftrightarrow{\alpha_i} \cdot \overrightarrow{E}$

Coulomb interaction

surface of the uracil molecule

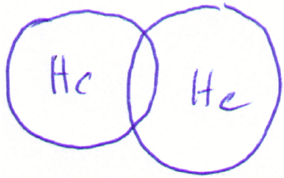




van der Waals interaction

Pauli repulsion

- electrons with the same spin avoid spatial overlap



- modeling:

$$E_{\text{ex}} = \exp\left[a - b \cdot R_{ij}\right]$$

$$E_{\text{ex}} = \left(\frac{\sigma}{R_{ij}}\right)^{12}$$

van der Waals interaction

dispersion due to correlation

- correlation – between electrons irrespective of spin, retained on longer distances
- instantaneous dipole → induced dipole → interaction

- orientation of dipoles is correlated – attractive interaction
- R^{-6}-dependence, proportional to polarizabilities

van der Waals interaction

most common function: Lennard-Jones 12-6 potential

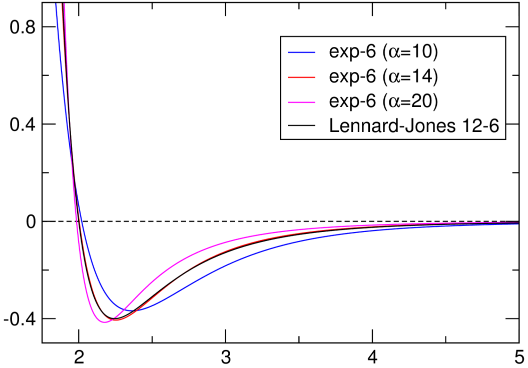
$$V(r) = 4\varepsilon\left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right)$$

- 2 parameters – σ and ε

repulsive: $\exp[-R]$ sometimes better than $R^{-12} \rightarrow$ exp-6 potential

- may be a better choice for phase transitions
- e.g. MM water would not freeze below 0 °C with LJ 12-6

van der Waals interaction



van der Waals interaction

parametrization

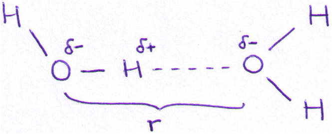
- challenging task in general
- fitting of params to experimental / quantum-chem. data
- e.g. relation to density and heat of vaporization
  – obvious in organic liquids – major interaction

Hydrogen bonding

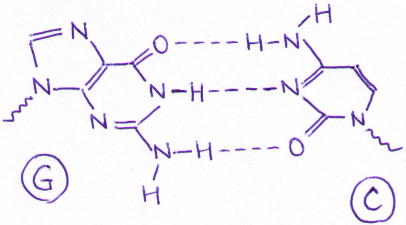
- crucial interaction in biomolecules
- interplay of several kinds of interactions
- typical binding energies: 20 kJ/mol
  higher for strongly polarized or even charged molecules
  or if there are several H-bonds (nucleobase pairs)
- early force fields – special potential functions for H-bonding
- modern force fields – no special treatment

Hydrogen bonding

H_2O dimer



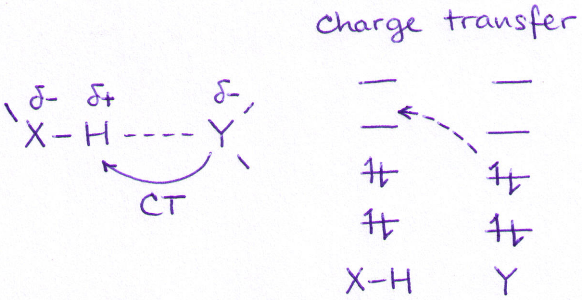
guanine:cytosine base pair



Hydrogen bonding

- Coulomb interaction is dominant
- vdW interaction
  - may become important, especially in weakly bound systems
  - crucial e.g. for angular dependence in $H_2CO \ldots H_2O$ etc.
- charge transfer contribution
  - cannot be covered by force fields due to constant charges
  - may be included in other terms effectively

Hydrogen bonding



Parametrization of bonded interactions

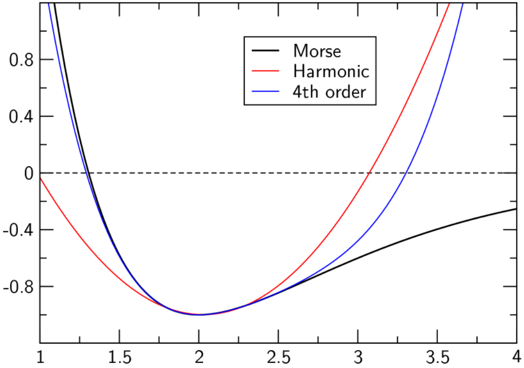
Bonds

- harmonic approximation (Taylor expansion up to 2nd order)

$$E(r) = \frac{1}{2}k(r - r_0)^2$$

  - 2 parameters – equilibrium distance and force constant
  - works in a narrow interval of distances
  - often sufficient (vibrations are within the interval)
- if bonds are to be created or broken (chemistry)
  another solution has to be sought
  → probably leave molecular mechanics ☹

Parametrization of bonded interactions



- for accurate vibration frequencies
  – 4th-order terms can be important to describe the curvature

Parametrization of bonded interactions

Angles

- harmonic approximation for the angle deformation

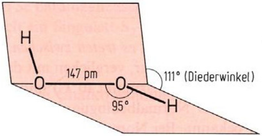
$$E_{\text{bend}}(\vartheta) = \frac{1}{2} k_{\vartheta} (\vartheta - \vartheta_0)^2$$

- 2 parameters needed – equilibrium angle and force constant
- from experiment (vib-rot spectra) or quantum chemistry

Parametrization of bonded interactions

Dihedral angles

- describe the rotation around covalent bonds
- defined by 4 atoms



Parametrization of bonded interactions

Dihedral angles

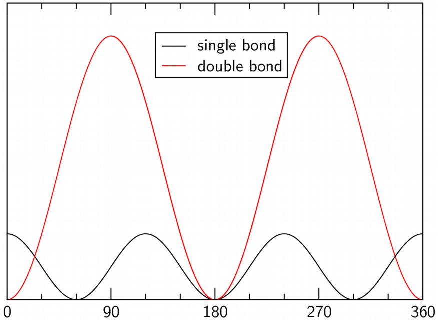
- describe the rotation around covalent bonds
- defined by 4 atoms
- potential energy – periodic function of the dihedral angle:

$$E(\omega) = \sum_{n=1,2,3,4,6} V_n \cos\left[n \cdot \omega - \gamma_n\right]$$

- V_n – amplitude (barrier), n – periodicity, γ – phase shift

Parametrization of bonded interactions

Dihedral angles – example: C–C single and C=C double bonds



The complete equation

$$E(R^N) =$$

$$\begin{aligned} &= \frac{1}{2}\sum_i k_i(r_i - r_i^0)^2 + \frac{1}{2}\sum_j k_j^\vartheta(\vartheta_j - \vartheta_j^0)^2 + \frac{1}{2}\sum_n V_n \cdot \cos\left[n\omega - \gamma_n\right] \\ &+ \sum_i^N \sum_{j=i+1}^N \left\{4\varepsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right) + \frac{1}{4\pi\varepsilon_0}\frac{q_i q_j}{r_{ij}}\right\} \end{aligned}$$

Molecular dynamics simulation

how to get things moving

Equations of motion

$$m \cdot \ddot{r} = F$$

ordinary differential equations of 2nd order

- have to be solved numerically
- solution proceeds in discreet steps of length Δt
- numerical integration starts at time t_0,
  where the initial conditions are specified
  – the positions r_0 and the velocities v_0
- calculations of forces at r_0 to get accelerations a_0
- then, an integrator calculates r and v at time $t_0 + \Delta t$
- accelerations → step → accelerations → step → . . .

Verlet integration method

a virtual step in positive time and in 'negative' time,
and Taylor expansion up to 2nd order:

$$\begin{aligned} r(t+\Delta t) &= r(t) + \dot{r}(t)\cdot\Delta t + \frac{1}{2}\ddot{r}(t)\cdot\Delta t^2 \\ r(t-\Delta t) &= r(t) - \dot{r}(t)\cdot\Delta t + \frac{1}{2}\ddot{r}(t)\cdot\Delta t^2 \end{aligned}$$

Verlet integration method

a virtual step in positive time and in ‘negative’ time,
and Taylor expansion up to 2nd order:

$$\begin{aligned} r(t+\Delta t) &= r(t) + \dot{r}(t)\cdot\Delta t + \frac{1}{2}\ddot{r}(t)\cdot\Delta t^2 \\ r(t-\Delta t) &= r(t) - \dot{r}(t)\cdot\Delta t + \frac{1}{2}\ddot{r}(t)\cdot\Delta t^2 \end{aligned}$$

add both equations – eliminate the velocity $\dot{r}$:

$$\begin{aligned} r(t+\Delta t) &= 2\cdot r(t) - r(t-\Delta t) + \ddot{r}(t)\Delta t^2 \\ \ddot{r}(t) &= a(t) = \frac{F(t)}{m} = -\frac{1}{m}\frac{\partial V}{\partial r}(t) \end{aligned}$$

Verlet integration method

another but equivalent formulation – velocity Verlet

$$\begin{aligned} r(t+\Delta t) &= r(t) + v(t)\cdot\Delta t + \tfrac{1}{2}a(t)\cdot\Delta t^2 \\ v(t+\Delta t) &= v(t) + \tfrac{1}{2}\left(a(t) + a(t+\Delta t)\right)\cdot\Delta t \end{aligned}$$

yet another – Leap-frog

$$\begin{aligned} v(t+\tfrac{1}{2}\Delta t) &= v(t-\tfrac{1}{2}\Delta t) + a(t)\cdot\Delta t \\ r(t+\Delta t) &= r(t) + v(t+\tfrac{1}{2}\Delta t)\cdot\Delta t \end{aligned}$$

Δt – crucial parameter

Let us say: we want to obtain a trajectory over a time interval T

- we perform M steps
- we have to evaluate the forces on atoms $M = T/\Delta t$ times

Computational cost of the calculation of forces

- major computational effort
- determines how many steps we can afford to make

Δt – crucial parameter

- we neglect contributions in Δt^3 and higher orders
  $\rightarrow$ error per step in the order of Δt^3
- keep the step short → make the error small
  but need too many steps to simulate certain time T
- make the step long → cut computational cost
  but increase the error and decrease stability
- compromise needed

Δt – crucial parameter

- fastest motion – hydrogen atoms, period around 10 fs
- rule of thumb – stable integration with $\Delta t \leq \frac{1}{10}$ fastest period
- practically, Δt of 1 fs is used (2 fs with special treatment)

- 1M calculations of forces needes for a trajectory of 1 ns
- large systems – multi-ns simulations routinely, μs possible

Introducing thermodynamics

what you simulate is what you would measure

Intro

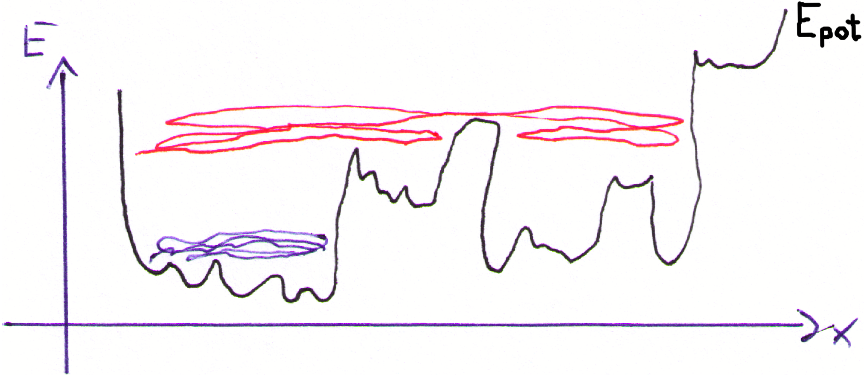
Solution of equations of motion – conserves total / internal energy

what we need – to control basic simulation parameters
– temperature and possibly pressure

significance of temperature
- determines which structures of the system are accessible
- different dynamics at high and at low temperatures

Intro

high E – multiple different structural ‘classes’ are reached
low E – restricted available structures



difference $E - E_{\text{pot}}$ corresponds to E_{kin} and temperature

Isolated system

- exchanges with surroundings neither energy (heat / work) nor matter (particles)
- total energy of system: $E = E_{\text{kin}} + E_{\text{pot}} = \text{const}$
- individually, E_{kin} and E_{pot} fluctuate in the course of time as they are being transformed into each other
- is what we get when using the Verlet method for a molecule

kinetic theory of gases → relation of E_{kin} and temperature:

$$\begin{aligned} \langle E_{\text{kin}} \rangle &= \frac{3}{2} NkT \\ \text{where } \langle E_{\text{kin}} \rangle &= \frac{1}{2} \sum_i m_i \left\langle v_i^2 \right\rangle \end{aligned}$$

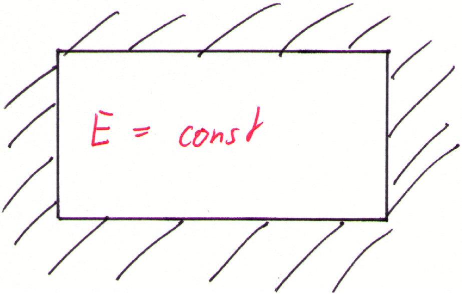
'local' T – fluctuates in time; may differ between parts of system

Isolated and closed system

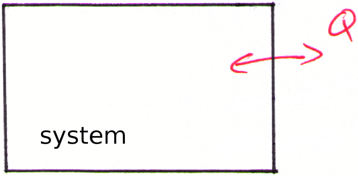
experimental setup (a test tube with a sample)

- usually in thermodynamic equilibrium with the surroundings
- temperature (and opt. pressure) equal as that of surr.

isolated system



closed system



system

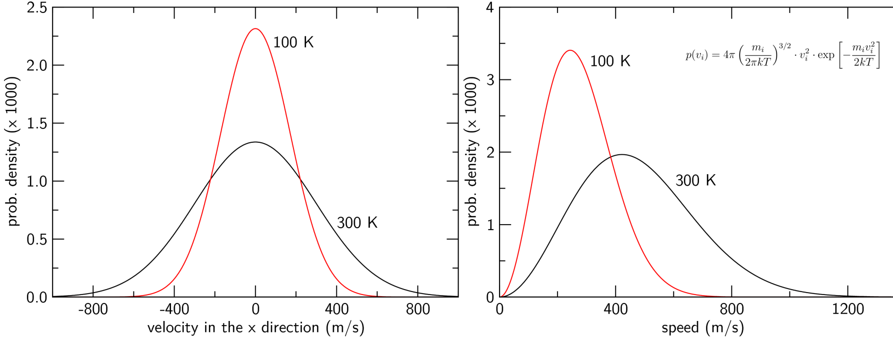
environment (bath)

Closed system

- thermal contact of system with surroundings
- exchange of energy in the form of heat
  until the temperature of surroundings is reached
- canonical ensemble
- velocity / speed of atoms – Maxwell–Boltzmann distribution

Canonical ensemble

Maxwell–Boltzmann distribution of velocity / speed (N_2, IG)



Naïve thermostat – scaling of velocities

in a Verlet MD simulation – 'instantaneous temperature' T deviates from the target T_{ref} (of bath = the surroundings)

$$T(t) = \frac{2}{3}\frac{E_{\text{kin}}(t)}{Nk} \neq T_{\text{ref}}$$

$T(t)$ – another name for E_{kin} determined by velocities
simple idea – scale the velocities by a certain factor λ:

$$\begin{aligned} T_{\text{ref}} &= \frac{1}{\frac{3}{2}Nk} \cdot \frac{1}{2} \sum_i m_i \left(\lambda \cdot v_i\right)^2 = \\ &= \lambda^2 \cdot \frac{1}{\frac{3}{2}Nk} \cdot \frac{1}{2} \sum_i m_i v_i^2 = \lambda^2 \cdot T \end{aligned}$$

Naïve thermostat – scaling of velocities

scaling of all velocities by $\lambda = \sqrt{T_\text{ref}/T} \;\rightarrow\; T_\text{ref}$ reached exactly

- rescaling the velocities affects the 'natural' way of evolution of the system
- velocities – not sure if the distribution is correct (M–B)
- importantly, system does not sample any canonical ensemble – very important because everything is calculated as averages:

$$\langle A\rangle = \frac{1}{Z}\int \rho\cdot A\,\mathrm{d}\vec{r}\,\mathrm{d}\vec{p}$$

- possibly: wrong sampling → wrong averages

Berendsen thermostat

How to avoid the drastic changes to the dynamics?

adjust velocities more smoothly, in the direction of T_{ref}

- temperature changes between two time steps according to

$$\Delta T = \frac{\Delta t}{\tau} \left(T_{\text{ref}} - T \right)$$

- rate of change of T (due to the change of velocities) is proportional to the deviation of actual T from T_{ref}
- constant of proportionality – relaxation time τ

Berendsen thermostat

- velocities are scaled by λ:

$$\begin{aligned} T_\text{new} &= T + \Delta T = T + \frac{\Delta t}{\tau}\left(T_\text{ref} - T\right) \\ \lambda &= \sqrt{\frac{T_\text{new}}{T}} = \sqrt{1 + \frac{\Delta t}{\tau}\left(\frac{T_\text{ref}}{T} - 1\right)} \end{aligned}$$

- usually: $\tau = 0.1 - 10$ ps
- T will fluctuate around the desired value T_ref
- problem – still does not generate correct canonical ensemble

Nosé–Hoover thermostat

- represents rigorously the canonical ensemble → ideal choice
- conceptionally and mathematically > difficult to understand
- heat bath is treated not as an external element
  rather as an integral part of the system
  is assigned an additional DOF s with fictitious mass Q
- eqns of motion for this extended system $(3N + 1$ DOF$)$:

$$\begin{aligned} \ddot{r}_i &= \frac{F_i}{m_i} - s \cdot \dot{r}_i \\ \dot{s} &= \frac{1}{Q}\left(T - T_{\text{ref}}\right) \end{aligned}$$

Temperature and thermostats

fluctuation of temperature – desired property
for canonical ensemble – variance of 'inst. temperature' T:

$$\sigma_T^2 = \left\langle (T - \langle T \rangle)^2 \right\rangle = \left\langle T^2 \right\rangle - \langle T \rangle^2$$

and relative variance

$$\frac{\sigma_T^2}{\langle T \rangle^2} = \frac{2}{3N}$$

large number of atoms N: fluctuations → 0
finite-sized systems: visible fluctuation of temperature
– is correct (feature of the canonical ensemble)

Introducing pressure

chemical reality – constant pressure rather than constant volume

goal – implement such conditions in simulations, too

How to calculate pressure? – first, calculate virial of force

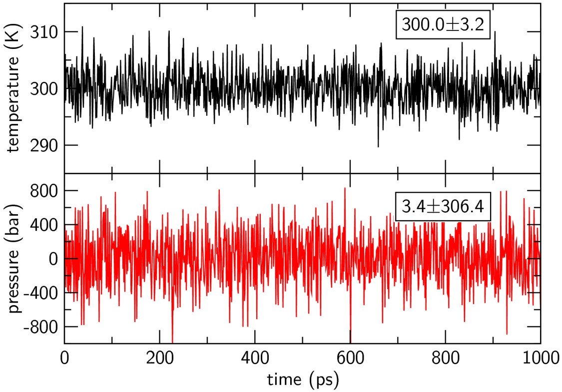
$$\Xi = -\frac{1}{2} \sum_{i<j} \vec{r}_{ij} \cdot \vec{F}_{ij}$$

($\vec{r}_{ij}$ distance of atoms i and j, $\vec{F}_{ij}$ – force between them)

$$P = \frac{2}{3V} \cdot (E_{\text{kin}} - \Xi) = \frac{2}{3V} \cdot \left(\frac{1}{2} \sum_i m_i \cdot |\vec{v}_i|^2 + \frac{1}{2} \sum_{i<j} \vec{r}_{ij} \cdot \vec{F}_{ij} \right)$$

Measuring pressure

T and P in an NPT simulation of a DNA oligomer in water
($T_{\mathrm{ref}} = 300$ K, $P_{\mathrm{ref}} = 1.0$ bar)



Controlling pressure

we can calculate the pressure
– so how do we maintain it at a constant value?
barostat – algorithm that is equivalent of a thermostat,
just that it varies volume of the box instead of velocities

alternatives are available:

- Berendsen barostat
  – direct rescaling of box volume
  – system coupled to a 'force / pressure bath' – piston
- Parrinello–Rahman barostat
  – extended-ensemble simulation
  – additional DOF for the piston