Statistical Mechanics of Proteins

Ioan Kosztin

Department of Physics & Astronomy
University of Missouri - Columbia

- **Equilibrium and non-equilibrium properties of proteins**
  - **Free diffusion of proteins**
- Coherent motion in proteins: temperature echoes
- Simulated cooling of proteins

Molecular Modeling

1. Model building
2. Molecular Dynamics Simulation
3. **Analysis** of the
   - model
   - results of the simulation

Collection of MD Data

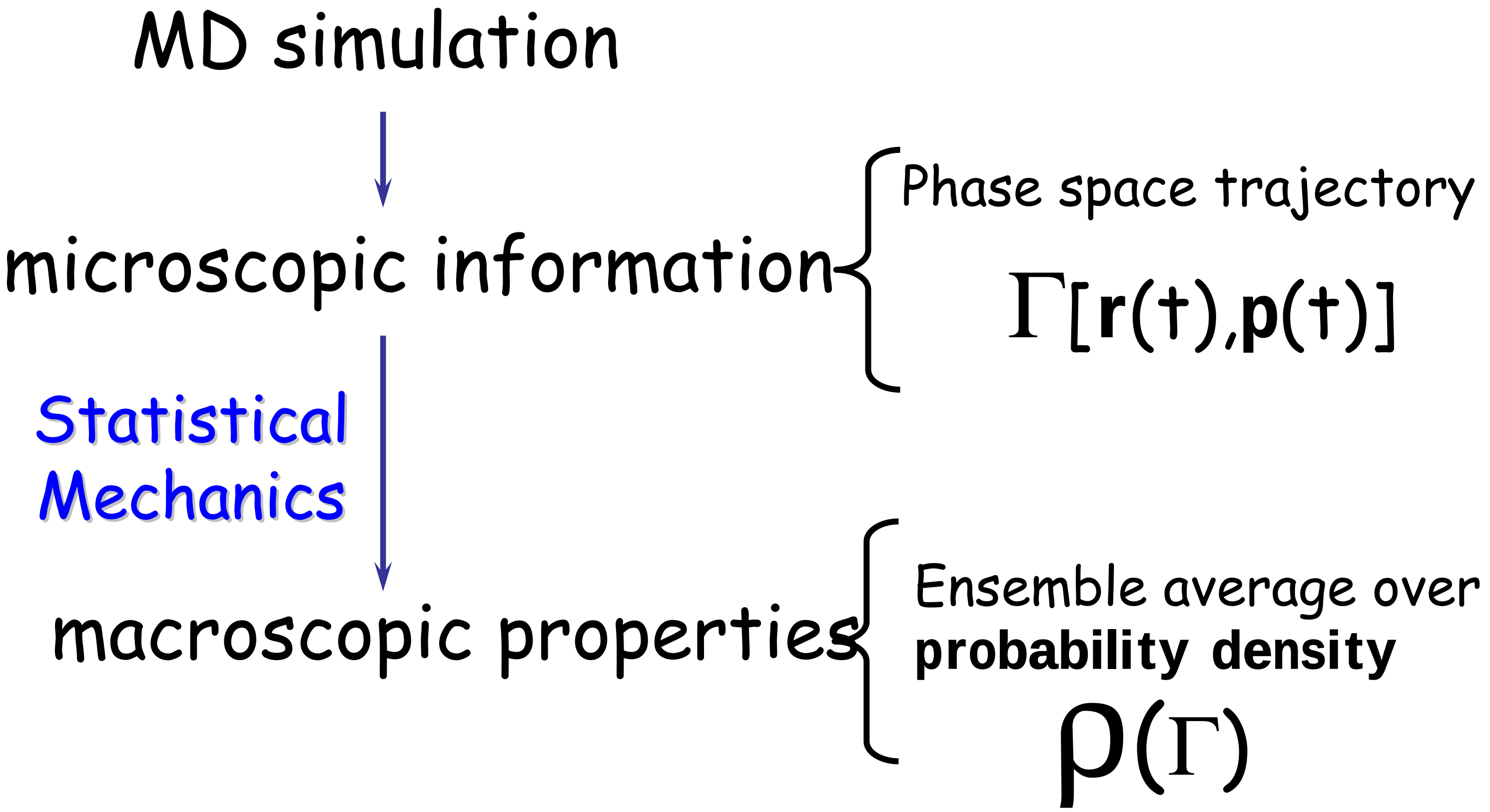
- DCD trajectory file
  - coordinates for each atom
  - velocities for each atom
- Output file
  - global energies
  - temperature, pressure, …

Analysis of MD Data

1. Structural properties
2. Equilibrium properties
3. Non-equilibrium properties

Can be studied via both **equilibrium** and **non-equilibrium** MD simulations

Equilibrium (Thermodynamic) Properties



MD simulation

microscopic information

Phase space trajectory

$\Gamma[\mathbf{r}(t),\mathbf{p}(t)]$

Statistical Mechanics

macroscopic properties

Ensemble average over **probability density**

$\rho(\Gamma)$

Statistical Ensemble

Collection of *large* number of replicas (on a macroscopic level) of the system

Each replica is characterized by the same macroscopic parameters (e.g., NVT, NPT)

The microscopic state of each replica (at a given time) is determined by Γ in phase space

Time vs Ensemble Average

For $t \to \infty$, $\Gamma(t)$ generates an ensemble with

$$\rho(\Gamma)d\Gamma = \lim_{t\to\infty} d\tau / t$$

Ergodic Hypothesis: **Time** and **Ensemble** averages are equivalent, i.e., $\langle A(r,p)\rangle_t = \langle A(\Gamma)\rangle_\rho$

Time average: $$\langle A\rangle_t = \frac{1}{T}\int_0^T dt\, A[\mathbf{r}(t),\mathbf{p}(t)]$$

Ensemble average: $$\langle A\rangle = \int d\Gamma\,\rho(\Gamma)\,A(\Gamma)$$

Thermodynamic Properties from MD Simulations

Thermodynamic (equilibrium) averages can be calculated via time averaging of MD simulation time series

$$\langle A \rangle \approx \frac{1}{N} \sum_{i=1}^{N} A(t_i)$$

Thermodynamic average

MD simulation time series

Finite simulation time means incomplete sampling!

Common Statistical Ensembles

1. **Microcanonical** (N,V,E):

$$\rho_{NVE}(\Gamma) \propto \delta[H(\Gamma) - E]$$

← Newton's eq. of motion

2. **Canonical** (N,V,T):

$$\rho_{NVT}(\Gamma) = \exp\{[F - H(\Gamma)]/k_B T\}$$

← Langevin dynamics

3. **Isothermal-isobaric** (N,p,T)

$$\rho_{NPT}(\Gamma) = \exp\{[G - H(\Gamma)]/k_B T\}$$

← Nose-Hoover method

Different simulation protocols [$\Gamma(t) \rightarrow \Gamma(t+\delta t)$] sample different statistical ensembles

Examples of Thermodynamic Observables

- Energies (kinetic, potential, internal,…)
- Temperature [*equipartition theorem*]
- Pressure [*virial theorem*]

Thermodynamic derivatives are related to mean square fluctuations of thermodynamic quantities

- Specific heat capacity C_V and C_P
- Thermal expansion coefficient α_P
- Isothermal compressibility β_T
- Thermal pressure coefficient γ_V

Mean Energies

Total (internal) energy:	$E = \frac{1}{N}\sum_{i=1}^{N} E(t_i)$	**TOTAL**
Kinetic energy:	$K = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{\boldsymbol{p}_j^2(t_i)}{2m_j}$	**KINETIC**
Potential energy:	$U = E - K$	**BOND** **ANGLE** **DIHED** **IMPRP** **ELECT** **VDW**

Note: You can conveniently use **namdplot** to graph the time evolution of different energy terms (as well as T, P, V) during simulation

Temperature

From the equipartition theorem $\langle p_k \partial H/\partial p_k \rangle = k_B T$

$$T = \frac{2}{3Nk_B}\langle K \rangle$$

Instantaneous *kinetic temperature*

$$\mathrm{T} = \frac{2K}{3Nk_B}$$

```
namdplot TEMP vs TS …
```

<u>Note</u>: in the NVT**P** ensemble $N \rightarrow N-N_c$, with $N_c=3$

Pressure

From the virial theorem $\langle r_k \partial H / \partial r_k \rangle = k_B T$

$$PV = N k_B T + \langle W \rangle$$

The *virial* is defined as

$$W = \frac{1}{3} \sum_{j=1}^{M} \boldsymbol{r}_j \cdot \boldsymbol{f}_j = -\frac{1}{3} \sum_{i,j>i} w(r_{ij})$$

pairwise interaction

with $w(r) = r\, dv(r) / dr$

Instantaneous *pressure* function (not unique!)

$$\mathrm{P} = \rho k_B T + W / V$$

Thermodynamic Fluctuations (TF)

$$\langle \delta A \rangle \approx \frac{1}{N} \sum_{i=1}^{N} \left[A(t_i) - \langle A \rangle \right]$$

Mean Square Fluctuations (MSF)

$$\langle \delta A^2 \rangle = \langle (A - \langle A \rangle)^2 \rangle = \langle A^2 \rangle - \langle A \rangle^2$$

According to *Statistical Mechanics*, the probability distribution of thermodynamic fluctuations is

$$\rho_{fluct} \propto \exp\left(\frac{\delta P \cdot \delta V - \delta T \cdot \delta S}{2 k_B T} \right)$$

TF in NVT Ensemble

In MD simulations distinction must be made between properly defined mechanical quantities (e.g., energy *E*, kinetic temperature T, instantaneous pressure P) and thermodynamic quantities, e.g., T, P, …

For example: $\langle \delta E^2 \rangle = \langle \delta \mathrm{H}^2 \rangle = k_B T^2 C_V$ ✓

But: $\langle \delta \mathrm{P}^2 \rangle \neq \langle \delta P^2 \rangle = k_B T / V \beta_T$ ✗

Other useful formulas: $\langle \delta K^2 \rangle = \dfrac{3N}{2} (k_B T)^2$

$$\langle \delta U^2 \rangle = k_B T^2 (C_V - 3N k_B / 2)$$

$$\langle \delta U \, \delta \mathrm{P} \rangle = k_B T^2 (\gamma_V - \rho k_B)$$

$$C_V = (\partial E / \partial T)_V$$

$$\gamma_V = (\partial P / \partial T)_V$$

How to Calculate C_V?

1. From definition

$$C_V = (\partial E / \partial T)_V$$

Perform multiple simulations to determine $E \equiv \langle E \rangle$ as a function of *T*, then calculate the derivative of $E(T)$ with respect to *T*

2. From the MSF of the total energy *E*

$$C_V = \langle \delta E^2 \rangle / k_B T^2$$

with $\langle \delta E^2 \rangle = \langle E^2 \rangle - \langle E \rangle^2$

Analysis of MD Data

1. Structural properties
2. Equilibrium properties
3. Non-equilibrium properties

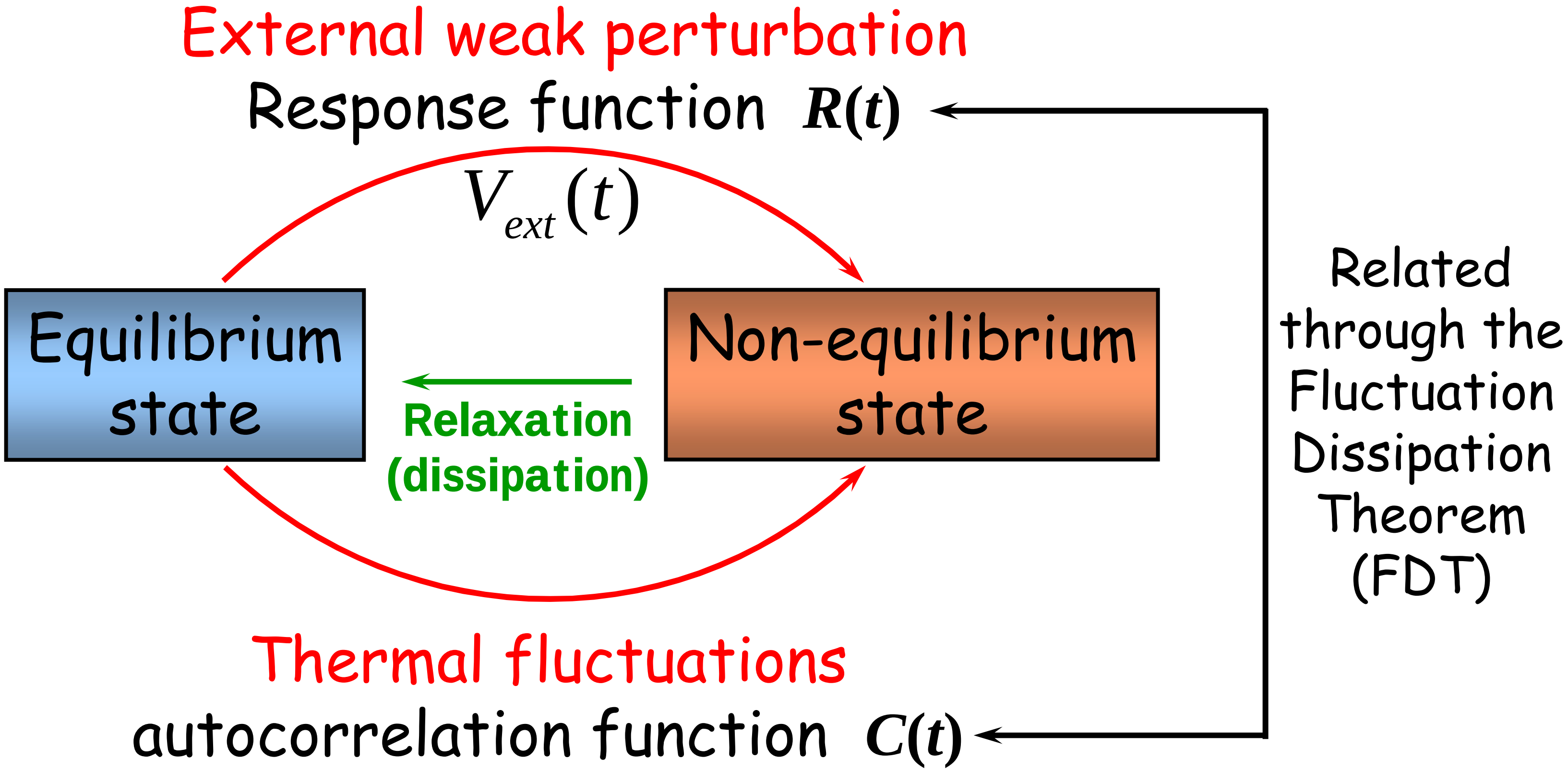
Can be studied via both **equilibrium** and/or **non-equilibrium** MD simulations

Non-equilibrium Properties

1. Transport properties
2. Spectral properties

Can be obtained from *equilibrium* MD simulations by employing *linear response theory*

Linear Response Theory



External weak perturbation

Response function $R(t)$

$V_{ext}(t)$

Equilibrium state

Non-equilibrium state

Relaxation (dissipation)

Thermal fluctuations

autocorrelation function $C(t)$

Related through the Fluctuation Dissipation Theorem (FDT)

Time Correlation Functions

$$C_{AB}(t-t') = \underbrace{\langle A(t)\,B(t')\rangle = \langle A(t-t')\,B(0)\rangle}_{\text{since } \rho_{eq} \text{ is } t \text{ independent !}}$$

$$\left.\begin{array}{ll} A \neq B & \text{cross-} \\ A = B & \text{auto-} \end{array}\right\} \text{correlation function}$$

Correlation time: $\tau_c = \int_0^\infty dt\, C_{AA}(t) / C_{AA}(0)$

Estimates how long the “memory” of the system lasts

In many cases (but not always): $C(t) = C(0)\exp(-t/\tau_c)$

Response Function

or *generalized susceptibility*

External perturbation: $V_{ext}(t) = -A \cdot f_{ext}(t)$

Response of the system: $\langle A(t) \rangle = \int_0^t dt' \, R(t-t') \, f_{ext}(t')$

Response function: $R(t) = \langle \{A(t), A\}_{PB} \rangle = -\beta \langle \partial_t A(t) \, A \rangle$

with $\beta = 1/k_B T$

Generalized susceptibility: $\chi(\omega) \equiv R(\omega) = \int_0^\infty dt \, e^{i\omega t} \, R(t)$

Rate of energy dissipation/absorption:

$$Q_\omega \equiv \overline{\langle A(t) \rangle \frac{df}{dt}} = \frac{1}{2} \omega \chi''(\omega) \, | f_0 |^2, \quad f(t) = \mathrm{Re} \, f_0 \, e^{-i\omega t}$$

Fluctuation - Dissipation Theorem

Relates $R(t)$ and $C(t)$, namely:

$$\chi''(\omega) = (\beta\omega/2)\,C(\omega)$$

In the static limit ($t \to \infty$): $C(0) = \langle A^2 \rangle = k_B T\, R(0)$

Note: quantum corrections are important when $k_B T \leq \hbar\omega$

$$\chi''(\omega) = \hbar^{-1} \tanh(\beta\hbar\omega/2)\,C(\omega)$$

Diffusion Coefficient

Generic transport coefficient: $\gamma = \int_0^\infty dt \, \langle \partial_t A(t) \, \partial_t A(0) \rangle$

Einstein relation: $2\gamma t = \langle [A(t) - A(0)]^2 \rangle$

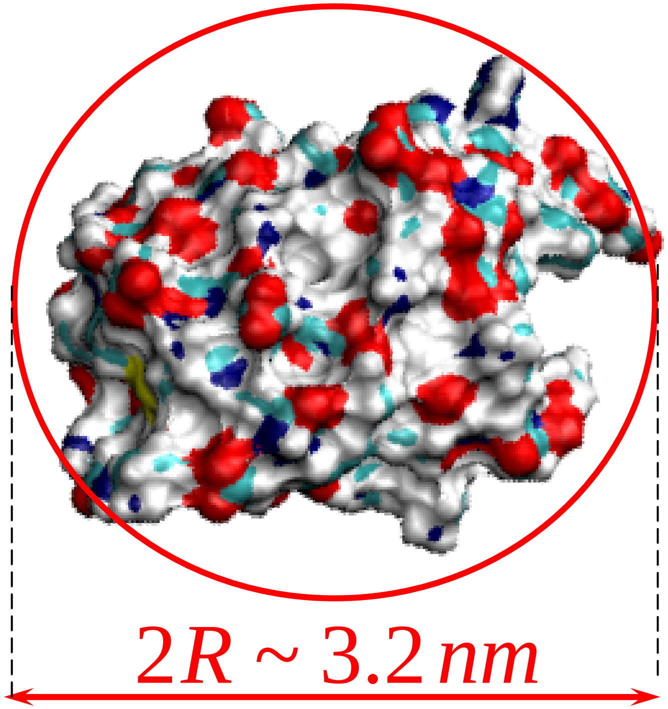
<u>Example</u>: *self-diffusion coefficient*

$$D = \frac{1}{3} \int_0^\infty dt \, \langle \mathbf{v}(t) \, \mathbf{v}(0) \rangle$$

$$6Dt = \left\langle [\mathbf{r}(t) - \mathbf{r}(0)]^2 \right\rangle$$

Free Diffusion (Brownian Motion) of Proteins

- in living organisms proteins exist and function in a viscous environment, subject to stochastic (random) thermal forces
- the motion of a globular protein in a viscous aqueous solution is diffusive



- e.g., *ubiquitin* can be modeled as a spherical particle of radius R~1.6nm and mass M=6.4kDa=$1.1\text{x}10^{-23}$ kg

Free Diffusion of Ubiquitin in Water

- ubiquitin in water is subject to two forces:
  - friction (viscous drag) force:

friction (damping) coeff

$$F_f = -\gamma \mathbf{v} \qquad \gamma = 6\pi\eta R$$

(Stokes law)

viscosity

  - stochastic thermal (Langevin) force:

$$F_L = \xi(t) \qquad \langle \xi(t) \rangle = 0$$

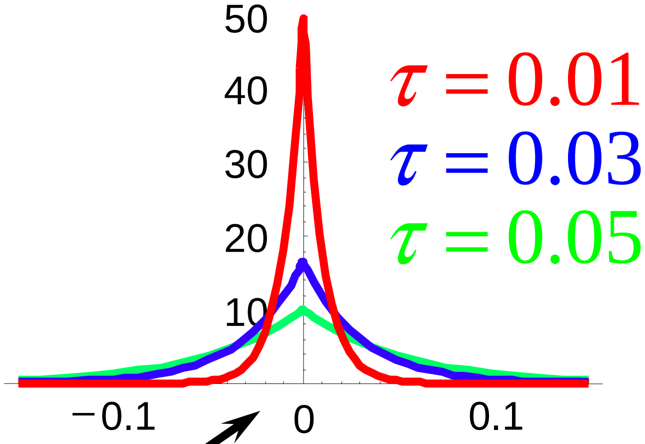
often modeled as a "**Gaussian white noise**"

$$\langle \xi(t)\,\xi(0) \rangle = 2\gamma k_B T\,\delta(t)$$

$k_B = 1.38\times10^{-23}\,J/K$ (*Boltzmann constant*), $T = temperature$

The Dirac delta function

$$\delta(t)=\begin{cases}\infty & \mathit{for}\ t=0\\ 0 & \mathit{for}\ t\neq 0\end{cases}$$



In practice, it can be approximated as:

$$\delta(t)\approx\delta_\tau(t)=\frac{1}{2\tau}\exp\left(-t/\tau\right),\ \ as\ \tau\rightarrow 0$$

⇒ δ(t) describes τ=0 correlation time ("white noise") stochastic processes

Useful formulas:

$$f(t)=\int_0^t f(t')\,\delta(t-t')\,dt' \qquad\qquad \delta(a\,t)=\delta(t)/\left|a\right|$$

Equation of Motion and Solution

Newton's 2nd law:

$$Ma = F_f + F_L \;\Rightarrow\; M\frac{dv}{dt} = -\gamma v + \xi(t)$$

Formal solution (using the variation of *const.* method):

$$v(t) = v_0 e^{-t/\tau} + \frac{1}{M} e^{-t/\tau} \int_0^t \xi(t') e^{t'/\tau} dt'$$

$$x(t) = x_0 + v_0\tau\left(1 - e^{-t/\tau}\right) + \frac{\tau}{M} e^{-t/\tau} \int_0^t \xi(t')\left(1 - e^{(t'-t)/\tau}\right) dt'$$

$\tau = \dfrac{M}{\gamma}$ = velocity relaxation (persistence) time

The **motion is stochastic** and requires **statistical description** formulated in terms of ***averages & probability distributions***

Statistical Averages

$$\boxed{\langle \xi(t) \rangle = 0 \qquad \langle \xi(t)\,\xi(0) \rangle = 2\gamma k_B T\,\delta(t)}$$

Exponential relaxation of *x* and *v* with characteristic time τ

$$\langle v(t) \rangle = v_0 e^{-t/\tau} \to 0 \quad as \quad t \to \infty$$

$$\langle x(t) \rangle = x_0 + v_0\tau\left(1 - e^{-t/\tau}\right) \to x_0 + v_0\tau \quad as \quad t \to \infty$$

$$\langle v^2(t) \rangle = \langle v(t) \rangle^2 + \frac{D}{\tau}\left(1 - e^{-t/\tau}\right) \to \frac{D}{\tau} \quad as \quad t \to \infty$$

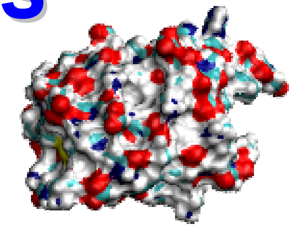
$$\langle x^2(t) \rangle = \langle x(t) \rangle^2 + 2Dt - 3D\tau + O\left(e^{-t/\tau}\right)$$

$$\Rightarrow \quad \Delta x(t) = \sqrt{\langle x^2 \rangle - \langle x \rangle^2} = \sqrt{2Dt} \quad as \;\; t \to \infty$$

Diffusion coefficient:
(*Einstein relation*)

$$D = k_B T / \gamma$$

Typical Numerical Estimates



example: ***ubiquitin*** - small globular protein

mass : $M \approx 8.6 kDa \approx 1.42\times10^{-23} kg$, size : $R \approx 1.6\, nm$,

density : $\rho = M/V \approx 10^3\, kg/m^3$, temperature : $T = 310\, K$

Property	Theory	Simulation
$v_T = \sqrt{3k_BT / M}$	$29.6\, m/s$	?
$\tau = M / \gamma$	$0.56\, ps$	?
$d = v_T\, \tau$	0.16 A	?
γ	$25.4\, pN \cdot s/m$	?
$D = \frac{k_BT}{\gamma} = \frac{1}{3} v_T^2 \tau$	$1.6\times10^{-10}\, m^2/s$	?
$\eta = \gamma / 6\pi R$	$0.9\, mPa \cdot s$	?

Thermal and Friction Forces

- Friction force:

$$F_f = \gamma v_T \approx 4.5\times 10^{-9}\,N = 4.5\,nN$$

- Thermal force:

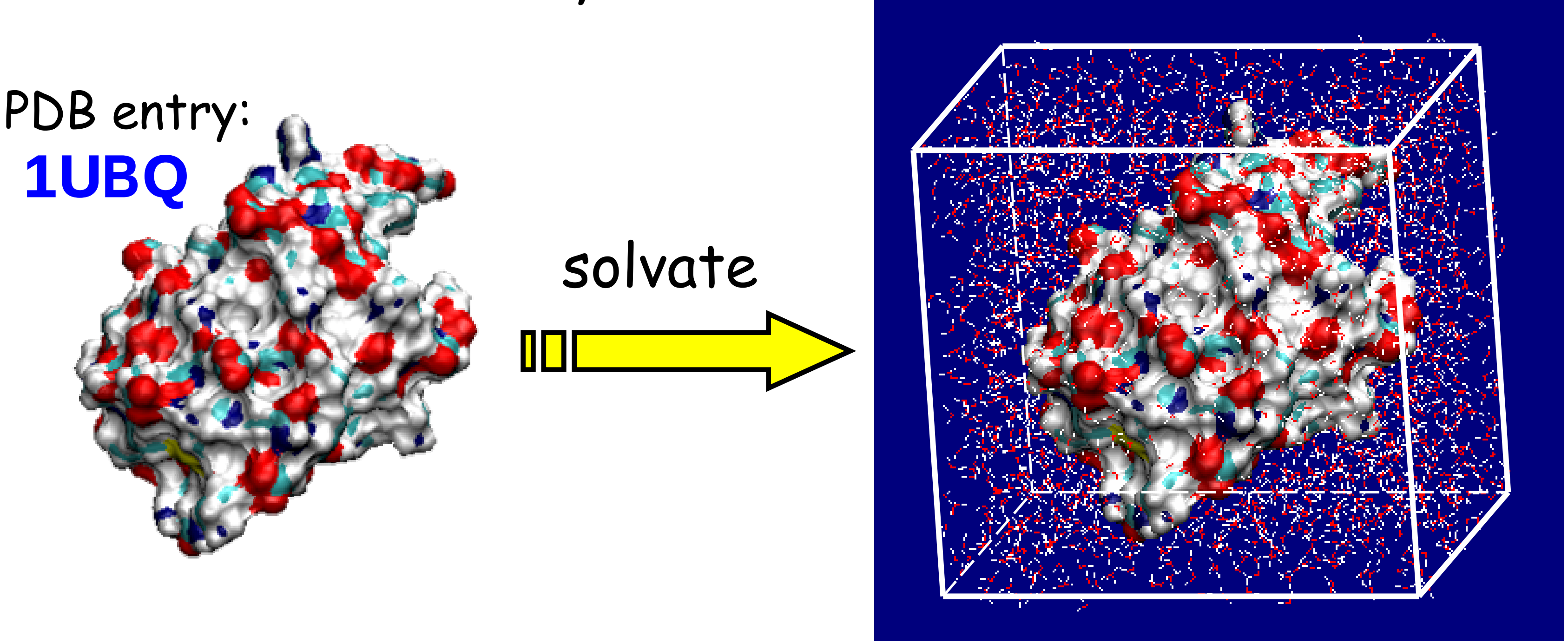
$$F_T = \sqrt{\frac{2\gamma k_B T}{\tau}} = \sqrt{\frac{2}{3}}\gamma v_T \sim F_f \approx 4.5\,nN$$

For comparison, the corresponding gravitational force:

$$F_g = Mg \sim 10^{-14}\,nN \ll F_f \sim F_T$$

Diffusion can be Studied by MD Simulations!

ubiquitin in water



total # of atoms: 7051 = 1231 (protein) + 5820 (water)

simulation conditions: NpT ensemble (T=310K, p=1atm), periodic BC, full electrostatics, time-step 2fs (SHAKE)

simulation output: Cartesian coordinates and velocities of all atoms saved at each time-step (10,000 frames = 40 ps) in separate DCD files

How To: vel.dcd —> vel.dat

- namd2 produces velocity trajectory (DCD) file if in the configuration file contains
```
velDCDfile  vel.dcd  ;# your file name
velDCDfreq  1        ;# write vel after 1 time-step
```
- load vel.dcd into VMD [e.g., `mol load psf ubq.psf dcd vel.dcd`]
note: run VMD in text mode, using the: -dispdev text option
- select only the protein (ubiquitin) with the VMD command
```
set ubq [atomselect top “protein”]
```
- source and run the following tcl procedure:
```
 source v_com_traj.tcl
v_com_traj COM_vel.dat
```
- the file “COM_vel.dat” contains 4 columns:
time [fs], v_x, v_y and v_z [m/s]
```
      70  12.6188434361 -18.6121653643 -34.7150913537
```

note: an ASCII data file with the trajectory of the COM coordinates can be obtained in a similar fashion

the v_COM_traj Tcl procedure

```
proc v_com_traj {filename {dt 2} {selection "protein"} {first_frame 0}
{frame_step 1} {mol top} args} {
    set outfile [open $filename "w"]
    set convFact  2035.4
    set sel [atomselect $mol $selection frame 0]
    set num_frames [molinfo $mol get numframes]
    for {set frame $first_frame} {$frame < $num_frames} {incr frame
$frame_step} {
        $sel frame $frame
        set vcom [vecscale $convFact [measure center $sel weight mass]]
        puts $outfile "$frame\t $vcom"
    }
    close $outfile
}
```

Goal: calculate D and τ

by fitting the theoretically calculated center of mass (COM) velocity autocorrelation function to the one obtained from the simulation

- **theory:** $$C_{vv}(t) = \left\langle v(t)\,v(0) \right\rangle = \left\langle v_0^2 \right\rangle e^{-t/\tau}$$

  $$\left\langle v_0^2 \right\rangle = \frac{k_B T}{M} = \frac{D}{\tau}$$ (equipartition theorem)

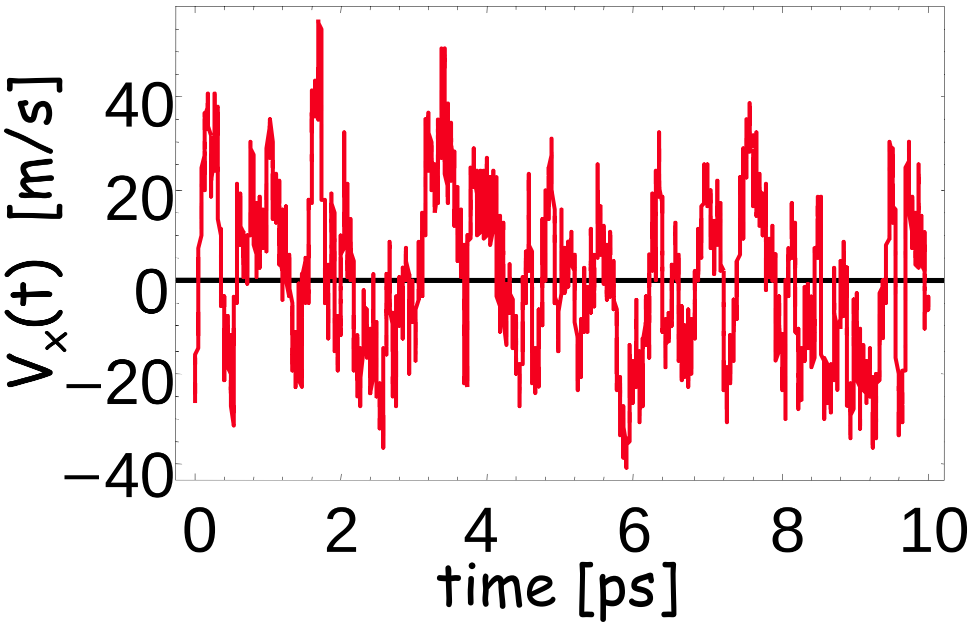
- **simulation:** consider only the x-component $\left(v_x \rightarrow v\right)$

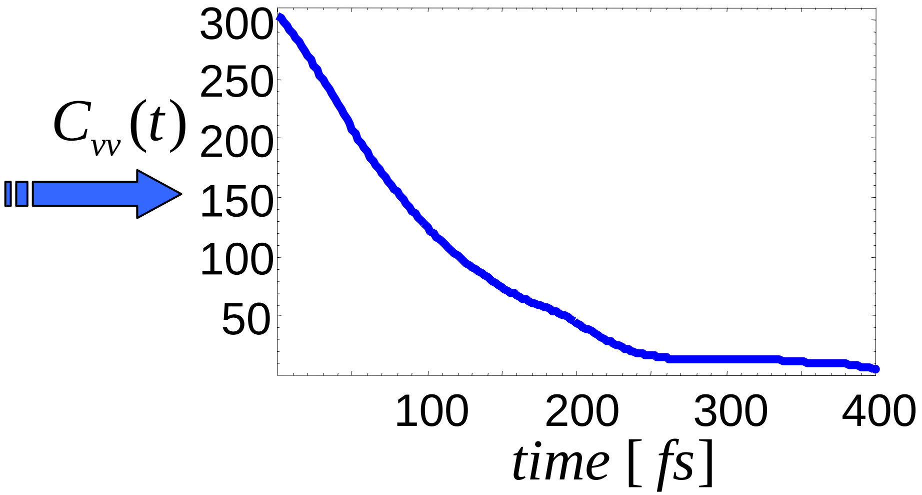
  replace ensemble average by time average

$$C_{vv}(t) \approx C_i = \frac{1}{N-i}\sum_{n=1}^{N-i} v_{n+i}\, v_n$$

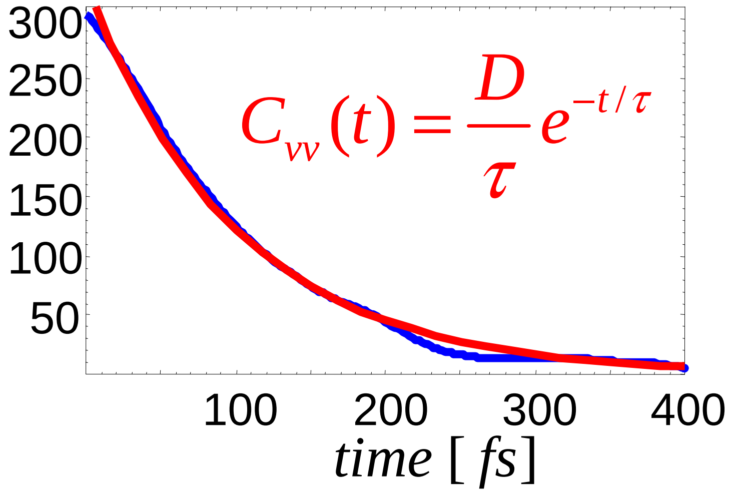
$t \equiv t_i = i\Delta t\,,\ v_n = v\left(t_n\right),\ \ N = \#$ of frames in vel.DCD

Velocity Autocorrelation Function





Fit

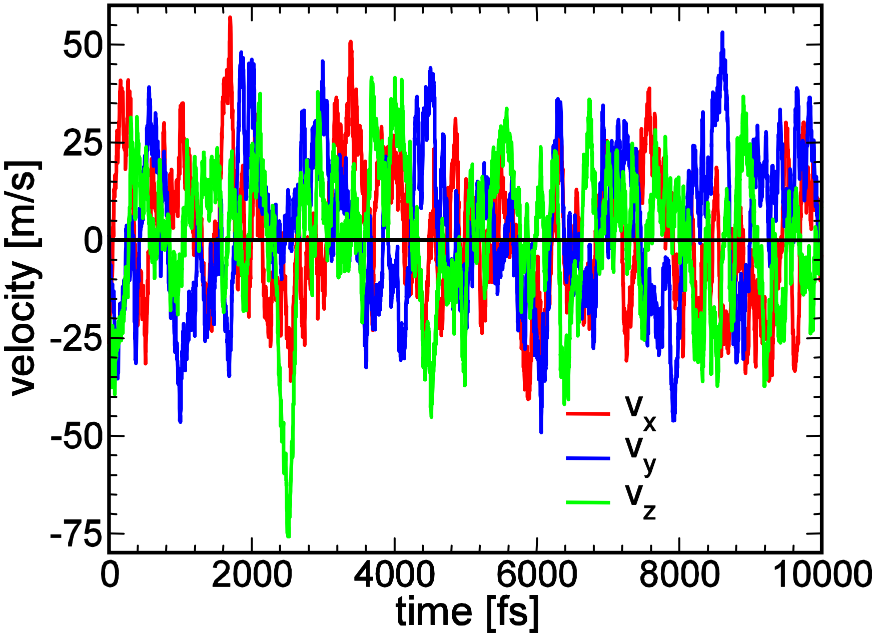


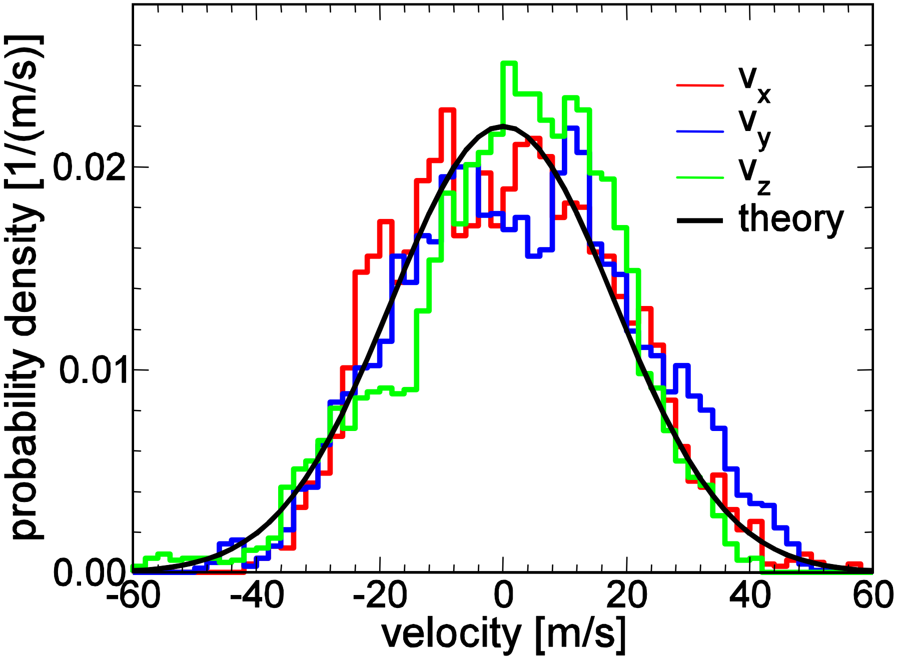
$$C_{vv}(t) = \frac{D}{\tau} e^{-t/\tau}$$

$$\tau \approx 0.1\, ps$$

$$D = k_B T/\gamma = \\ = \left\langle v_x^2 \right\rangle \tau \;\approx 3.3\times 10^{-11}\, m^2 s^{-1}$$

Probability distribution of $v_{x,y,z}$

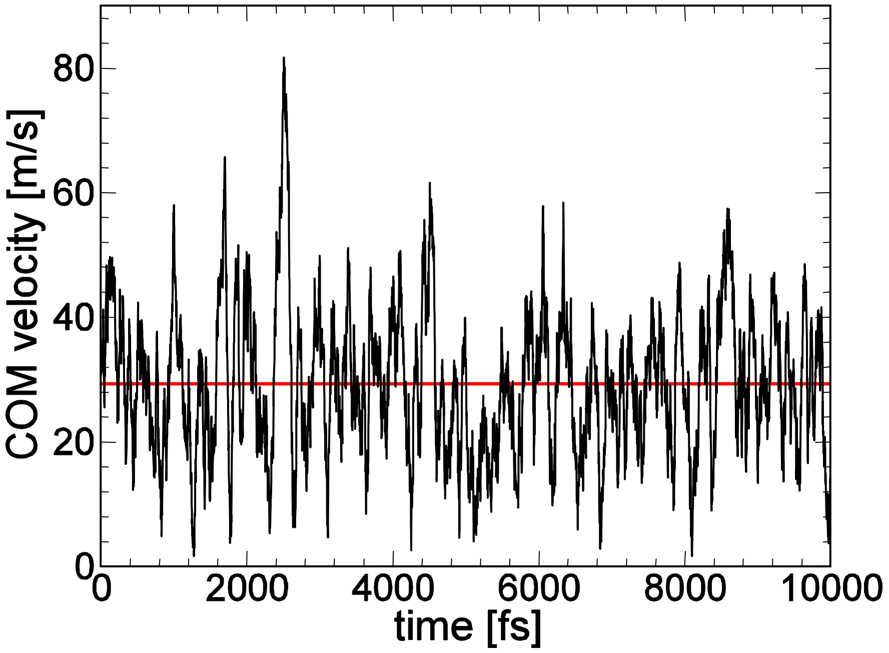


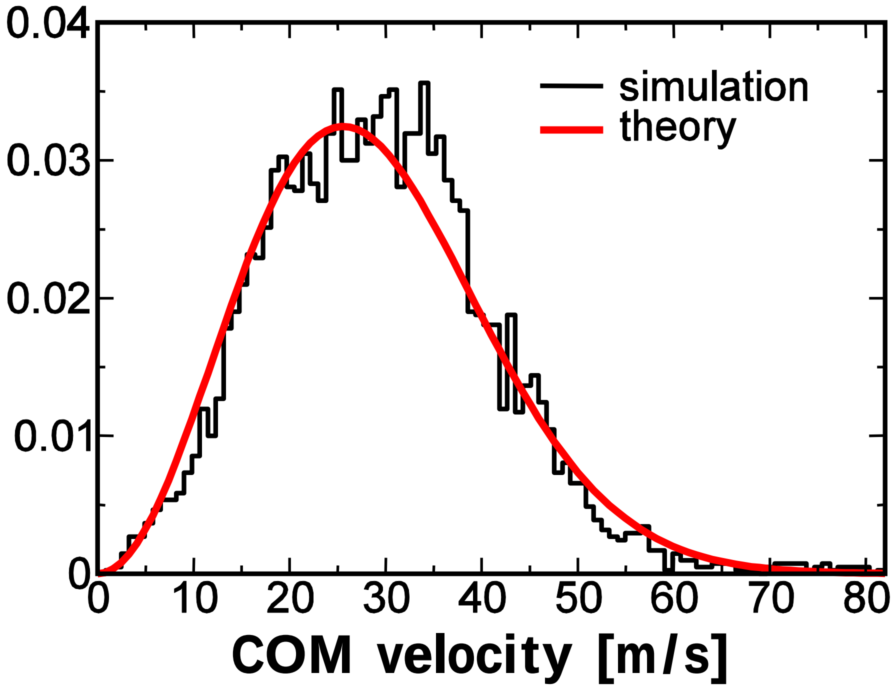


$$\boxed{\begin{aligned} p(v) &= \left(2\pi\left\langle v^2\right\rangle\right)^{-1/2}\exp\left(-v^2/2\left\langle v^2\right\rangle\right) \\ &= \sqrt{\tau/2\pi D}\,\exp\left(-\tau v^2/2D\right) \end{aligned}}$$

with $v \equiv v_{x,y,z}$

Maxwell distribution of v_{COM}





$$P(v)dv = p(v_x)\,p(v_y)\,p(v_z)dv_x dv_y dv_z$$

$$= (\tau/2\pi D)^{3/2}\exp\left(-\tau v^2/2D\right)4\pi v^2 dv$$

$$= \sqrt{\frac{2}{\pi}}\left(\frac{M}{k_B T}\right)^{3/2} v^2 \exp\left(-\frac{Mv^2}{2k_B T}\right)dv$$

What have we learned ?

soluble, globular proteins in aqueous solution at physiological temperature execute free diffusion (Brownian motion with typical parameter values:

Property	Theory	Simulation
$v_T = \sqrt{3k_B T / M}$	$29.6\, m/s$	$31.6\, m/s$
$\tau = M/\gamma$	$0.56\ ps$	$0.1\ ps$
$d = v_T\ \tau$	$0.16\,A$	$0.03\,A$
γ	$25.4\ pN \cdot s/m$	$141.6\ pN \cdot s/m$
$D = \frac{k_B T}{\gamma} = \frac{1}{3} v_T^2 \tau$	$1.6 \times 10^{-10}\ m^2/s$	$0.3 \times 10^{-10}\ m^2/s$
$\eta = \gamma/6\pi R$	$0.9\ mPa \cdot s$	$4.7\ mPa \cdot s$

How about the motion of parts of the protein ?

- parts of a protein (e.g., side groups, a group of amino acids, secondary structure elements, protein domains, …), besides the **viscous**, **thermal forces** are also subject to a **resultant force** from the rest of the protein
- for an effective degree of freedom x (reaction coordinate) the equation of motion is

$$m\ddot{x} = -\gamma\dot{x} + f(x) + \xi(t)$$

In the harmonic approximation $f(x) \approx -kx$

and we have a 1D **Brownian oscillator**