Molecular computer experiment

Also *simulation* or *pseudoexperiment*

REAL EXPERIMENT	COMPUTER EXPERIMENT
Record everything in a lab notebook	Record everything in a lab notebook
Choose method (device, assay)	Choose method (MD, MC, . . .)
Build the experimental apparatus (from parts)	Download/buy/write a computer program (blocks of code)
Purchase chemicals, synthetise if not available	Get a force field, fit/calculate parameters if not available (e.g., partial charges)
Prepare the experiment	Prepare initial configurations, etc.
Perform the experiment, watch what's going on	Run the code, observe time development, control quantities, etc.
Analyse and calculate	Calculate mean values (with error estimates)
Clean the laboratory	Make backups, erase temporary files

MD or MC?

Often, MC and MD can be applied to similar systems.

MD

- realistic models, complex molecules (bonds, angles. . .)
- condensed matter in general (fluids, solutions; biochemistry)
- kinetic quantities (diffusivity, viscosity. . .)
- better parallelization, more packages available

MC

- simple qualitative models (lattice, hard-sphere-like)
- dilute systems
- critical phenomena
- fluid equilibria
- overcoming barriers, exchange of molecules, etc. is easier with MC
- less efficient parallelization, fewer packages available

Is it correct?

Systematic errors:

- inaccurate molecular model (force field)
- neglected quantum effects, neglected many-body forces . . .
- small sample (finite-size effects)
- insufficient time scale (long correlations, bottleneck problems)
- method problems: integration errors (too long timestep), inappropriate thermostat/barostat, not equilibrated enough, inaccurate treatment of Coulomb forces. . .

Random (stochastic, statistical) errors are essential in stochastic methods

- time-correlated
- can be decreased by long calculations

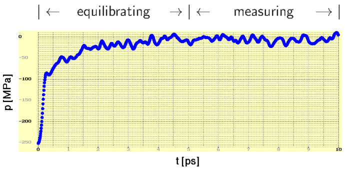
Uncertainty (in metrology) includes critical assessment of both the systematic and random errors

Warning: different terminology in different fields (mathematical statistics, metrology, physics, chemistry)

Simulation methodology

[sleep 3;simul/spceE.sh]

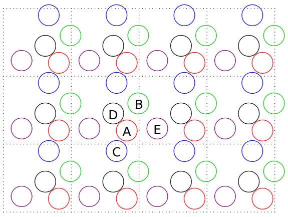
- Start (initial configuration):
  - experimental structure (biomolecules)
  - crystal → liquid (melt), gas → liquid (shrink); Packmol
  - random configuration (overlaps of molecules = problem in MD) problem for "ill-defined" models (TIP4P etc.)
  - lattice models: crystal/chaos
  - MD: velocities = Maxwell–Boltzmann (approximation enough)
- Equilibration → watch graphically (convergence/time profile)
- Measuring the quantities of interest incl. estimates of errors



Boundary conditions

[simolant]

- free (vacuum) – droplet, protein in a vacuum . . .
  1000 molecules in a cube 10^3 ⟶ $8^3 = 512$ are "inside"
- periodic (cyclic, torroidal)



- walls (hard, soft, smoothed, made of atoms), pores, slab, . . .

Periodic boundary conditions: MD

```
REAL L edge size of the cubic simulation box (cell)
VECTOR r1, r2 where vector r = (r.x,r.y,r.z)
              both vectors must lie in the basic box
VECTOR dr := r2 - r1 difference of vectors
                     (ignoring the boundary conditions)

IF      dr.x < -L/2 THEN dr.x := dr.x + L
ELSE IF dr.x >  L/2 THEN dr.x := dr.x - L

IF      dr.y < -L/2 THEN dr.y := dr.y + L
ELSE IF dr.y >  L/2 THEN dr.y := dr.y - L

IF      dr.z < -L/2 THEN dr.z := dr.z + L
ELSE IF dr.z > -L/2 THEN dr.z := dr.z - L
Vector dr now goes from r1 to the nearest image of r2

Squared distance to the nearest image:
REAL rr := dr.x**2 + dr.y**2 + dr.z**2
```

Periodic boundary conditions: MC

In MC, usually the vector $\vec{r}_{12} = \mathrm{r2} - \mathrm{r1}$ is not needed, the distance is enough

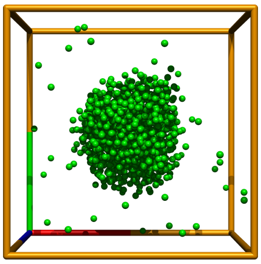
```
REAL L edge size of the cubic simulation box (cell)
VECTOR r1, r2 where vector r = (r.x,r.y,r.z)
              both vectors must lie in the basic box
VECTOR dr := r2 - r1 difference of vectors
                     (ignoring the boundary conditions)

REAL rr := (L/2 - abs(L/2-abs(dr.x)))**2
         + (L/2 - abs(L/2-abs(dr.y)))**2
         + (L/2 - abs(L/2-abs(dr.z)))**2
```

Calculations

[../simul/ar/showdrop.sh]

Example. We simulate an argon droplet in a periodic cubic simulation cell. Let us have $N = 1000$ atoms and temperature 85 K. The distance between surfaces of periodic images of droplets should be equal to the droplet diameter. Calculate the size of the box in Å. Argon density is $\rho = 1.4\,\mathrm{g\,cm^{-3}}$, molar mass $M(\mathrm{Ar}) = 40$ g/mol.



molar volume: $V_\mathrm{m} = M/\rho$

volume per 1 atom: $V_1 = V_\mathrm{m}/N_\mathrm{A}$

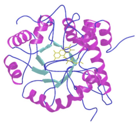
volume of N atoms: $V = NV_1 = NM/\rho N_\mathrm{A}$
$= 1000 \cdot 0.040\,\mathrm{kg\,mol^{-1}}/(1400\,\mathrm{kg\,m^{-3}} \cdot 6.022\times10^{23}\,\mathrm{mol^{-1}})$
$= 4.744\times10^{-27}\,\mathrm{m^3}$

sphere radius: $\frac{4}{3}\pi R^3 = V \Rightarrow R = 2.24\times10^{-9}$ m

box size: $L = 90$ Å

One more example

Example. Consider a globular protein of molecular weight of 20 kDa. The density of the protein is $1.35\,\mathrm{g\,cm^{-3}}$. Calculate the approximate protein diameter.



$$m = \frac{20\,\mathrm{kg\,mol^{-1}}}{6.022\times10^{23}\,\mathrm{mol^{-1}}} = 3.32\times10^{-23}\,\mathrm{kg}$$

or 1 Da = $1\,\mathrm{g\,mol^{-1}}/N_\mathrm{A} = 1.6605\times10^{-27}\,\mathrm{kg}$ (atomic mass unit)

$$m = 20000 \times 1.6605\times10^{-27}\,\mathrm{kg} = 3.32\times10^{-23}\,\mathrm{kg}$$

$$V = \frac{m}{\rho} = \frac{3.32\times10^{-23}\,\mathrm{kg}}{1350\,\mathrm{kg\,m^{-3}}} = 2.46\times10^{-26}\,\mathrm{m^3}$$

$$\frac{4\pi}{3}r^3 = \frac{\pi}{6}d^3 = V$$

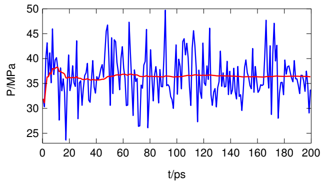
$$d = \sqrt[3]{\frac{6V}{\pi}} = \sqrt[3]{\frac{6\cdot2.46\times10^{-26}\,\mathrm{m^3}}{\pi}} = 3.61\times10^{-9}\,\mathrm{m} \doteq \underline{3.6\,\mathrm{nm}} = 36\,\text{Å}$$

Measurements

Trajectory = sequence of configurations (MD: in time)

Convergence profile:

- time development of a quantity (time profile, —) problems better seen
- cumulative (running average, —) can estimate the inaccuracy



Type of statistical treatment:

- averaged values (← ergodic hypothesis)
- less often fluctuations

Type of quantity:

- mechanical (temperature, pressure, internal energy, order parameters. . .)
- entropic (S, F, μ, . . .)
- structure (correlation functions, number of neighbors, analysis of clusters. . .)
- auxiliary or control quantities (order parameters, integrals of motion in MD)

Random errors

quantity = (estimate of the mean value) ± (estimate of the error)

Arithmetic average (example of a **statistic**, also *statistical functional, estimator*, in metrology *measurement function*):

$$\overline{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$$

statistic = estimator
statistics = field of mathematics

Standard error = standard deviation of the statistic, usually denoted as σ

$$\sigma_X = \sqrt{\left\langle \left(\overline{X} - \langle X\rangle\right)^2 \right\rangle}$$

For **uncorrelated** (independent) X_i and large m, $\overline{X}$ has Gaussian distribution

The estimate of the standard error of uncorrelated data:

$$\sigma_X^{\text{estim}} = \sqrt{\frac{\sum_{i=1}^{m} \Delta X_i^2}{m(m-1)}}, \qquad \text{where } \Delta X_i = X_i - \overline{X}$$

$\langle X\rangle \in$	probability
$(\overline{X}-\sigma_X, \overline{X}+\sigma_X)$	67.3%
$(\overline{X}-2\sigma_X, \overline{X}+2\sigma_X)$	95.45%
$(\overline{X}-1.96\sigma_X, \overline{X}+1.96\sigma_X)$	95.00%
$(\overline{X}-3\sigma_X, \overline{X}+3\sigma_X)$	99.730%
$(\overline{X}-5\sigma_X, \overline{X}+5\sigma_X)$	99.9999427%

Customs and terminology

"estimate of quantity with error/uncertainty": $123.4 \pm 0.5 \equiv 123.4(5) \equiv 123.4_5$

What exactly is the "error/uncertainty"?

Physics:

"error/uncertainty" = σ_X

- (Estimated) standard error/uncertainty, standard deviation of the average (or other statistic)
- Loosely: error/uncertainty, standard deviation, error margin/bar, . . .
- Custom certainty level in physics = $5\sigma_X$ ($\langle X\rangle \in \overline{X} \pm \sigma_X$ with probability 99.999943%)

Biology, economy, politology, engineering, pharmacology:

"error/uncertainty" = $2\sigma_X$ more precisely: $1.96\sigma_X$

- Interval of confidence at 95% confidence level
- Loosely: Interval of confidence (without specifying the confidence level)

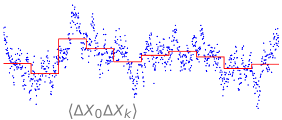
Chemistry: mostly ignored, if given, nobody knows whether error/uncertainty = σ_X or $2\sigma_X$

The type of error/uncertainty must be specified

Analysis of time series and error estimation

[cd simul; autoregr.sh 200+2000]

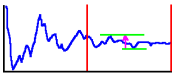
Problem: correlations



- block method: $\overline{X}_j = \frac{1}{B}\sum_{i=1}^{B} X_{i+(j-1)B}$
- analysis of correlations ⇒

$$\sigma_X = \sqrt{\frac{\sum_{i=1}^{m} \Delta X_i^2}{m(m-1)}(1+2\tau)} \qquad \tau = \sum_{k=1}^{\infty} c_k \qquad c_k = \frac{\langle \Delta X_0 \Delta X_k\rangle}{\langle (\Delta X)^2\rangle}$$

MC: c_k is monotonously decreasing [ex.: $c_k = \sum_{\lambda\neq 1} c_\lambda \lambda^k$, $\lambda \in (-1,1)$]
MD: $c_k \to c(t)$ (time autocorrelation function): damped oscillations



- even better = both approaches combined:
  first to block a bit, then $\tau \approx c_1$
- from running average (roughly ≈ 10 blocks):

$$\sigma_X^{\text{estim}} \approx 0.6[\max_{\text{2nd half}}(X) - \min_{\text{2nd half}}(X)]$$

or to be on the safe side (this formula is approximate):

$$\text{err}_X \approx \max_{\text{2nd half}}(X) - \min_{\text{2nd half}}(X)$$

⇒ $\langle X\rangle \in (\overline{X} - \text{err}_X, \overline{X} + \text{err}_X)$ with probability ≈ 85% (for long enough time series)

Exercise/Example

[simul/errplot2.sh 32+4096 0.9]

- Generate random correlated data (1st order process):

$$X_{k+1} = qX_k + u$$

where $u = u_{[0,1)}$ or u_{Gauss} etc., and $|q| < 1$.

- Calculate the arithmetic average incl. error by different methods

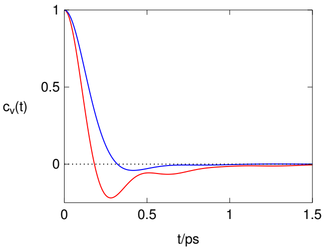
Note: it is known analytically,

$$\sigma_X = \sqrt{\frac{1+q}{1-q}}\sqrt{\frac{\operatorname{Var} X}{m}} = \frac{1}{1-q}\sqrt{\frac{\operatorname{Var} u}{m}}$$

where the variance, or fluctuation, is defined by $\operatorname{Var} X = \langle (X-\overline{X})^2\rangle$

Time autocorrelation function

Velocity-velocity autocorrelation function of liquid argon:
— 150 K, 1344 $\mathrm{kg\,m^{-3}}$,
— 120 K, 1680 $\mathrm{kg\,m^{-3}}$.
Results from a 100 ps trajectory for 216 Lennard-Jones particles



Typical behavior (MC + MD):

- fluid: $\lim_{t\to\infty} c(t) = \text{const}\, t^{-3/2}$ (hydrodynamic tail)
- jumps between states: $c(t) \propto \lambda^t$ (λ just below 1)

Error analysis – addition and subtraction

Sum of independent measurements: squares of standard deviations are additive

Example. Let us perform thermodynamic integration $I = \int_0^1 f(x)\mathrm{d}x$ approximately by the Simpson's formula:

$$I = \int_0^1 f(x)\mathrm{d}x \approx \frac{1}{6}[f(0) + 4f(0.5) + f(1)]$$

For $f(x)$ we have measured the following data with standard errors:

x	0	0.5	1
$f(x)$	1.34(5)	1.57(3)	1.77(6)

Calculate I including the error estimate.

$$I = \frac{1}{6}[1.34 + 4\times 1.57 + 1.77] = 1.565$$

$$\sigma(I)^2 = (0.05/6)^2 + (0.03\times 4/6)^2 + (0.06/6)^2 = 0.000569 \quad\Rightarrow\quad \sigma(I) = 0.024$$

$$I = 1.565(24)$$

Error analysis – division and multiplication

For division and multiplication, the squares of relative errors are additive

Example. Calculate 3.46(7)/0.934(13).

fraction: $3.46/0.934 = 3.704$

rel. error = $\sqrt{\left(\frac{0.07}{3.46}\right)^2 + \left(\frac{0.013}{0.934}\right)^2} = 0.0246$

abs. error = $3.704 \times 0.0246 = 0.091$

$\underline{3.46(7)/0.934(13) = 3.70(9)}$ (or rounded up: 3.70(10))

Error analysis

Error of function f of a variable with error is (linearized; i.e., for small σ):

$$f(x\pm\sigma_x) = f(x) \pm f'(x)\sigma_x$$

$$\ln(x\pm\sigma_x) = \ln x \pm \frac{\sigma_x}{x}, \qquad \exp(x\pm\sigma_x) = \exp x \pm \sigma_x \exp x, \qquad \frac{1}{x\pm\sigma_x} = \frac{1}{x} \pm \frac{\sigma_x}{|x|^2}$$

Example. Calculate the activity of $\mathrm{H^+}$ from pH = 2.125(5).

activity:

$$a_{\mathrm{H^+}} = 10^{-2.125} = \exp(-2.125\times\ln 10) = 0.00750$$

error Method 1:

$$\sigma = 0.005 \times \ln 10 \times a = 0.000086$$

error Method 2:

$$\sigma = |10^{-2.125} - 10^{-2.125-0.005}| = 0.000087$$

activity with error (uncertainty) estimate:

$$\underline{a_{\mathrm{H^+}} = 0.00750(9)}$$