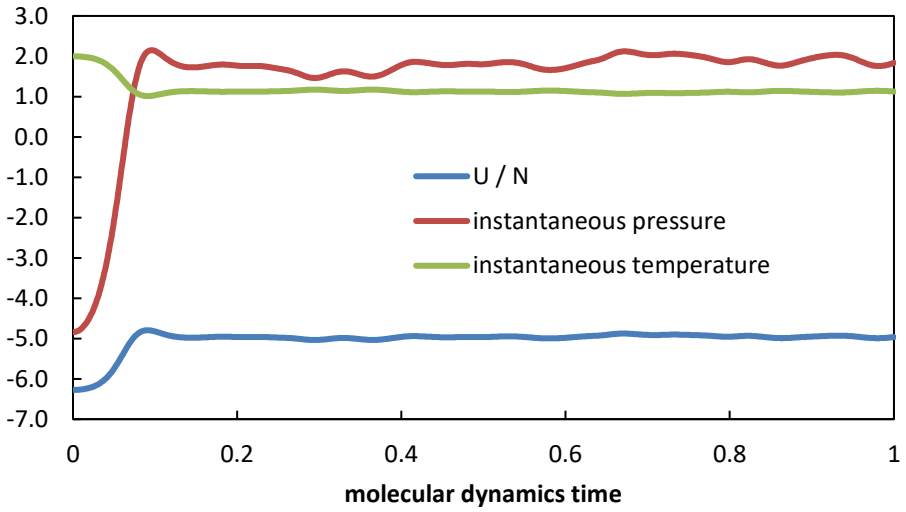
Today's lecture: how to compute thermodynamic properties like the temperature and pressure, and kinetic properties like the diffusivity and viscosity, from molecular dynamics and other simulations

Equilibration and production periods

Often we start our simulation with initial velocities and positions that are not representative of the state condition of interest (e.g., as specified by the temperature and density). As such, we must **equilibrate** our system by first running the simulation for an amount of time that lets it evolve to configurations representative of the target state conditions. Once we are sure we have equilibrated, we then perform a **production** period of simulation time that we used to study the system and/or compute properties at the target state conditions.

How do we know if we have well-equilibrated our system? One approach is to monitor the time-dependence of simple properties, like the potential energy or pressure. The following is taken from a 864-particle molecular dynamics simulation of the Lennard-Jones system. Initially, the atoms are placed on an fcc lattice and the velocities are sampled from a $T = 2.0$ (reduced units) distribution. The crystal melts to a liquid phase.



For the above system, an equilibration time might be ~0.2 time units. After equilibration, many quantities will still fluctuate—and *should* fluctuate if we are correctly reproducing the properties of the statistical mechanical ensemble of interest (here, the NVE ensemble).

At a basic level, we want the equilibration time to be at least as long as the **relaxation time** of our system, broadly defined here as the largest time scale for molecular motion.

One approach for estimating the relaxation time is to use diffusion coefficients or other measures of molecular motion. In a bulk liquid, for example, we might think of a relaxation time scale as that corresponding for one molecule to move a distance equal to one molecular diameter (σ). If we know the diffusion coefficient D, we can compute a relaxation time τ_{relax} from

$$\tau_{\text{relax}} \sim \frac{\sigma^2}{D}$$

Thus our equilibration time should at least exceed τ_{relax} several times over. Notice that D can vary with state conditions (e.g., temperature), and this should be taken into account if we perform multiple simulations at different conditions.

Simple estimators

What kinds of properties or **observables** can we compute from the production period of our simulation? The following discusses some variables commonly of interest. Each of these involves **averages** over the simulation duration.

Energies

The average kinetic and potential energies in our simulation are given by:

$$\langle K \rangle = \frac{1}{n}\sum K_i \qquad \langle U \rangle = \frac{1}{n}\sum U_i$$

where we sum n independent samples of the instantaneous kinetic and potential energies at different time points in the simulation. Remember, the statistical behavior of these sums shows that the error in our estimate goes as $n^{-1/2}$.

Temperature

There is no rigorous microscopic definition of the temperature in the microcanonical ensemble. Instead, we must use macroscopic thermodynamic results to make a connection here. Namely,

$$\frac{1}{T} = \left(\frac{\partial S}{\partial E}\right)_{V.N} = k_B \left(\frac{\partial \ln \Omega}{\partial E}\right)_{V.N}$$

It can be shown (using the equipartition theorem) that the average kinetic energy relates to the temperature via:

$$\langle K \rangle = n_{\text{DOF}} \frac{k_B T}{2}$$

Here, n_{DOF} is the number of degrees of freedom in the system. For a system of N atoms that conserves net momentum,

$$n_{\text{DOF}} = 3N - 3$$

However, for large enough systems the subtraction of the 3 center of mass degrees of freedom has little effect since it is small relative to $3N$. If rigid bonds are present in our system (treated in a later lecture), we also lose one degree of freedom per each.

Thus we can make a **kinetic estimate of the temperature:**

$$T = \frac{2\langle K \rangle}{k_B n_{\text{DOF}}}$$

Note that we can define, operationally, an **instantaneous kinetic temperature**:

$$T_{\text{inst}} = \frac{2K}{k_B n_{\text{DOF}}} \qquad T = \langle T_{\text{inst}} \rangle$$

Because K fluctuates during a simulation, T_{inst} also fluctuates. Note that this is an **estimator** of the temperature in that we must perform an average to compute it. The same ideas about independent samples also apply here.

Although it is not as frequently used, we can also compute a **configurational estimate of the temperature** [Butler, Ayton, Jepps, and Evans, J. Chem. Phys. 109, 6519 (1998)]:

$$k_B T_{\text{config}} = \left\langle \frac{\mathbf{f}^N \cdot \mathbf{f}^N}{-\nabla \cdot \mathbf{f}^N} \right\rangle$$

This estimate depends on the forces and their derivatives (via the denominator). Since the forces depend only on the atomic positions, and not momenta, this is termed a configurational estimate. We must also average over multiple configurations and correlation times in order to compute this temperature accurately. Both the kinetic and configurational temperatures are equal in the limit of infinite simulation time and equilibrium.

Velocity rescaling

Typically we want to perform a simulation at a specified temperature. For an NVE simulation, this means that we want to adjust the total energy such that the average temperature is equal to the one we specify. We can adjust the total energy easily by changing the momenta.

The most common approach is to **rescale the velocities** at periodic time intervals based on the deviation of the instantaneous temperature from our set point temperature. This is a form of a **thermostat.** If we rescale all of the velocities by:

$$\mathbf{v}^N_{\text{new}} = \lambda \mathbf{v}^N$$

We want:

$$T = \frac{\sum m_i \lambda^2 |\mathbf{v}_i|^2}{k_B n_{\text{DOF}}}$$

where T is our setpoint temperature. Solving for λ,

$$\lambda = \sqrt{\frac{k_B T n_{\text{DOF}}}{\sum m_i |\mathbf{v}_i|^2}}$$

Typically this rescaling is not done at every time step but only periodically (e.g., every 100-1000 time steps). Technically speaking, rescaling should be performed at a frequency related to the velocity autocorrelation time, discussed below.

One problem with velocity rescaling is that it affects the dynamics of the simulation run and is an artificial interruption to Newton's equations of motion. In particular, velocity rescaling means that the total energy E is no longer conserved, and that transport properties cannot be accurately computed. An alternative and perhaps better approach is:

1. First equilibrate the system using periodic velocity rescaling at the desired temperature.
2. Run a short production phase with velocity rescaling. Due to the rescaling, E will fluctuate. Compute an average total energy $\langle E \rangle$.
3. Turn off velocity rescaling.
4. Rescale the momenta such that the total energy equals $\langle E \rangle$. That is, given the current configuration with potential energy U, rescale the momenta and kinetic energy K to satisfy $K = \langle E \rangle - U$.
5. The simulation can then be evolved in time normally (NVE dynamics) and should average to the desired temperature, to within the errors in determining $\langle E \rangle$.

There are more sophisticated ways of performing temperature regulation but the above approach is perhaps the simplest. Moreover, this approach preserves the *true* NVE dynamics of the system, the only truly correct dynamics.

Pressure

To compute the pressure, we often use the **virial theorem**:

$$P = \frac{1}{3V}\left\langle 3Nk_BT + \sum \mathbf{f}_i \cdot \mathbf{r}_i \right\rangle$$
$$= \frac{1}{3V}\left\langle 2K + \sum \mathbf{f}_i \cdot \mathbf{r}_i \right\rangle$$

This expression is derived for the canonical ensemble (constant NVT), but it is often applied to molecular dynamics simulations regardless (NVE). For large enough systems, the difference between the two is very small.

The expression above is not generally used for systems of pairwise-interacting molecules subject to periodic boundary conditions. Instead, we can rewrite the force sum:

$$\sum_i \mathbf{f}_i \cdot \mathbf{r}_i = \sum_i \left(\sum_j \mathbf{f}_{ij} \right) \cdot \mathbf{r}_i$$
$$= \sum_{i,j} \mathbf{f}_{ij} \cdot \mathbf{r}_i$$
$$= \sum_{i<j} \mathbf{f}_{ij} \cdot \mathbf{r}_i + \sum_{i>j} \mathbf{f}_{ij} \cdot \mathbf{r}_i$$
$$= \sum_{i<j} \mathbf{f}_{ij} \cdot \mathbf{r}_i + \sum_{i<j} \mathbf{f}_{ji} \cdot \mathbf{r}_j$$
$$= \sum_{i<j} \mathbf{f}_{ij} \cdot \left(\mathbf{r}_i - \mathbf{r}_j \right)$$
$$= -\sum_{i<j} \frac{du(r_{ij})}{dr} r_{ij}$$

Thus,

$$W \equiv -\sum_{i<j} \frac{du(r_{ij})}{dr} r_{ij}$$

$$P = \frac{1}{3V}\langle 2K + W \rangle$$

Here, W is called the **virial.** Notice that W involves a sum of pairwise interactions. We therefore need to compute it in our pairwise loop, alongside the energies. Oftentimes, the calculations we use for the pairwise energies can be re-used in the loop. Take the Lennard-Jones system for example, in dimensionless units:

$$U = \sum_{i<j} 4\left(r_{ij}^{-12} - r_{ij}^{-6}\right)$$

$$W = \sum_{i<j} 24\left(2r_{ij}^{-12} - r_{ij}^{-6}\right)$$

For systems involving rigid bonds (discussed later), the forces acting to hold the bonds rigid must be computed and added to the overall virial.

Heat capacity

One way we might measure the heat capacity is to perform multiple simulations at different temperatures and then numerically estimate

$$C_V = \frac{dE}{dT} \approx \frac{E(T + \Delta T) - E(T)}{\Delta T}$$

Alternatively, we can estimate C_V from a single simulation using energy fluctuations. For the canonical ensemble (and approximately the microcanonical one), we can write:

$$\begin{aligned} C_V &= \frac{d(K+U)}{dT} \\ &= \frac{n_{\text{DOF}} k_B}{2} + \frac{dU}{dT} \\ &= \frac{n_{\text{DOF}} k_B}{2} + \frac{\langle U^2 \rangle - \langle U \rangle^2}{k_B T^2} \\ &= \frac{n_{\text{DOF}} k_B}{2} + \frac{\sigma_U^2}{k_B T^2} \end{aligned}$$

That is, we can measure the heat capacity from the variance in the potential energy. The last term in this equation is often termed the **configurational heat capacity**.

Other quantities

It is relatively easy to measure the **enthalpy**, which stems from a mechanical average:

$$H = U + PV$$

On the other hand, it is very challenging to compute entropic or free-energetic quantities, like S, A, G, μ. We will discuss advanced simulation approaches for determining these quantities later in the course. Unlike the quantities we have studied so far, free-energetic quantities require computation of **distributions** of simulation observables, not just averages of them.

Statistics of averages

Basic averages

Consider the computed average potential energy of a simulation. For a production period of n MD time steps, we could compute

$$\overline{U} = \frac{1}{n}\sum_{i=1}^{n} U_i$$

In this section, we will use an overbar to indicate an estimate deduced from a single, finite-duration simulation. It will be more informative for now if we neglect the discretized nature of our solutions to the dynamic trajectories and instead represent this as an integral:

$$\overline{U} = \frac{1}{t_{\text{tot}}}\int_0^{t_{\text{tot}}} U(t)dt$$

This expression isn't specific to the potential energy. For any observable A for which we want to compute the average,

$$\bar{A} = \frac{1}{t_{\text{tot}}}\int_0^{t_{\text{tot}}} A(t)dt$$

These averages of observables correspond to finite-duration simulations. There are two ways in which we might expect to see errors in our results:

- The simulation time is not long enough to reduce statistical error in $\bar{A}$. Only in the limit $t_{\text{tot}} \to \infty$ will we rigorously measure the true, statistical-mechanical average that we expect from thermodynamics. In practice, we really only need to take this integral to a moderate number of **correlation times** of the property A, which we discuss below.

- The simulation is not at equilibrium. In this case, we need to extend the equilibration period before computing this integral.

In what follows, we will use the following notational definitions. Let

$$\bar{A} = \frac{1}{t_{\text{tot}}}\int_0^{t_{\text{tot}}} A(t)dt \qquad \langle A\rangle = \lim_{t_{\text{tot}}\to\infty}\frac{1}{t_{\text{tot}}}\int_0^{t_{\text{tot}}} A(t)dt$$

That is, $\bar{A}$ denotes a simulation average, while $\langle A\rangle$ denotes the true statistical-mechanical equilibrium average for A that we would expect in the limit of infinite simulation time, in which our system is at equilibrium.

Correlation times

Assume we can perform a simulation that initially is fully equilibrated at the desired equilibrium conditions. If we were to perform multiple **trials** or **runs** of our simulation, we would get an estimate for $\bar{A}$ that would be different each time because of the finite length for which we perform them. We could obtain a number of measurements from different runs:

$$\bar{A}_1, \bar{A}_2, \bar{A}_3, \ldots$$

We want to know what the expected variance of $\bar{A}$ is, relative to the true value $\langle A \rangle$. This is the squared error in our measurement of the average using finite simulation times:

$$\sigma_{\bar{A}}^2 = \langle (\bar{A} - \langle A \rangle)^2 \rangle$$

Here, the brackets indicate an average over an infinite number of simulations we perform. We can simplify this expression:

$$\begin{aligned}
\sigma_{\bar{A}}^2 &= \langle \bar{A}^2 \rangle - \langle A \rangle^2 \\
&= \left\langle \frac{1}{t_{\text{tot}}^2} \left[\int_0^{t_{\text{tot}}} A(t) dt \right] \left[\int_0^{t_{\text{tot}}} A(t) dt \right] \right\rangle - \langle A \rangle^2 \\
&= \left\langle \frac{1}{t_{\text{tot}}^2} \int_0^{t_{\text{tot}}} \int_0^{t_{\text{tot}}} A(t) A(t') dt' \, dt \right\rangle - \langle A \rangle^2 \\
&= \frac{1}{t_{\text{tot}}^2} \int_0^{t_{\text{tot}}} \int_0^{t_{\text{tot}}} \langle A(t) A(t') \rangle dt' \, dt - \langle A \rangle^2
\end{aligned}$$

In the last line, we moved the average into the integrand.

In the integrals, we have a double summation of all $A(t)A(t')$ pairs at different time points. For two specific time points, $t = t_1$ and $t' = t_2$, the identical products $A(t_1)A(t_2)$ and $A(t_2)A(t_1)$ both appear as the integrand variables pass over them. This enables us to consider only the unique time point pairs of t, t' for which $t' < t$, multiplying by two:

$$\sigma_{\bar{A}}^2 = \frac{2}{t_{\text{tot}}^2} \int_0^{t_{\text{tot}}} \int_0^{t} \langle A(t) A(t') \rangle dt' \, dt - \langle A \rangle^2$$

We can also simplify things because Newton's equations are symmetric in time. First, the average

$$\langle A(t) A(t') \rangle$$

should not depend on the absolute value of the times, but only their relative value, because at equilibrium we can look at our simulation at any two relative points in time and we would expect to get the same average. Therefore, we can shift this average in time by $-t'$:

$$\sigma_{\bar{A}}^2 = \frac{2}{t_{\text{tot}}^2}\int_0^{t_{\text{tot}}}\int_0^t \langle A(t-t')A(0)\rangle dt'\,dt - \langle A\rangle^2$$

Since the simulations start at equilibrium, we have

$$\langle A\rangle = \langle A(0)\rangle$$

$$\langle A^2\rangle = \langle A(0)^2\rangle$$

$$\sigma_A^2 = \langle A^2\rangle - \langle A\rangle^2 = \langle A(0)^2\rangle - \langle A(0)\rangle^2$$

Notice that σ_A^2 (without overbar on the A) gives the equilibrium variance of A, or that expected from a single, long equilibrium simulation. It is different from $\sigma_{\bar{A}}^2$, which estimates the variance in the average $\bar{A}$, or the squared error in the average we compute from run to run. We expect σ_A^2 to be finite, constant, and characteristic of the equilibrium fluctuations, while we expect $\sigma_{\bar{A}}^2$ to approach zero as we make our simulations longer and longer.

With these ideas, we can rewrite this expression as:

$$\sigma_{\bar{A}}^2 = \frac{2\sigma_A^2}{t_{\text{tot}}^2}\left[\int_0^{t_{\text{tot}}}\int_0^t C_A(t-t')dt'\,dt\right]$$

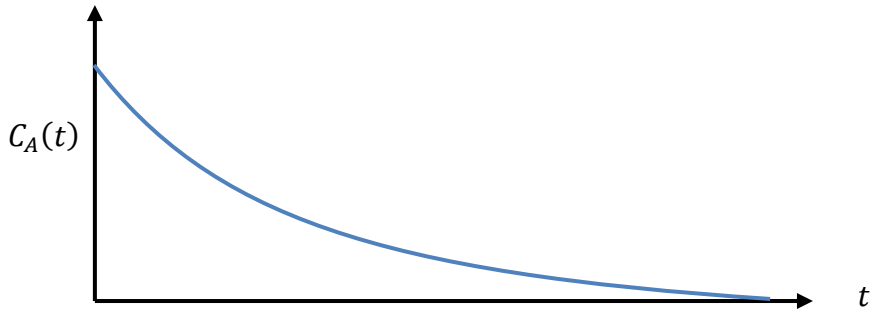
Here, C_A is the **time autocorrelation function** for A. Its formal definition is

$$C_A(t) \equiv \frac{\langle A(t)A(0)\rangle - \langle A(0)\rangle\langle A(0)\rangle}{\langle A(0)A(0)\rangle - \langle A(0)\rangle\langle A(0)\rangle} = \frac{\langle A(t)A(0)\rangle - \langle A(0)\rangle\langle A(0)\rangle}{\sigma_A^2}$$

Physically, it measures how correlated the variable A is at some time t with its value at initial time 0. By the definition above we see that

$$C_A(t=0) = 1 \qquad C_A(t\to\infty) = 0$$

Schematically, the correlation function may look something like this:



Autocorrelation functions decay with time, since at long times, a measurement is uncorrelated from its value at earlier times. We can define an **autocorrelation time** as:

$$\tau_A \equiv \int_0^\infty C_A(t)dt$$

If the total simulation length is longer than this time, $t_{\text{tot}} \gg \tau_A$, the expression for the variance in $\bar{A}$ can be rewritten approximately as:

$$\sigma_{\bar{A}}^2 \approx \frac{2\sigma_A^2}{t_{\text{tot}}^2}\left[\int_0^{t_{\text{tot}}} \tau_A dt\right]$$
$$= \frac{2\sigma_A^2}{t_{\text{tot}}/\tau_A}$$

We can define an effective number of **independent samples** n_A such that:

$$n_A \equiv t_{\text{tot}}/2\tau_A$$

$$\sigma_{\bar{A}}^2 = \frac{\sigma_A^2}{n_A}$$

This result is an extremely important one. It says several things:

- The squared error in any quantity for which we average in simulation decreases as one over the effective number of **independent samples**.
- Samples that we use in our average to compute $\bar{A}$ are only independent if we pick them to be spaced at least $2\tau_A$ units apart in time.
- We will not get better statistical accuracy by averaging the value of A for every single time step in our simulation. We get just as good accuracy by averaging the value of A for times spaced $2\tau_A$ units of time apart.

Block averaging

We want to make sure that we are including enough independent samples in our estimates of different property averages. A very basic approach would be to estimate the largest time scale in our system, the relaxation time, and make sure we perform the simulation for a large number of these times. This is perhaps the most common approach.

Alternatively, we could compute τ_A. Indeed, there are procedures for estimating correlation functions from simulations. We could perform a very long simulation, compute the correlation function, and estimate τ_A using the integral definition of it. However, it can be a significant effort to determine correlation functions in our simulations since they require long runs a priori.

Instead, we can use a simple **block averaging** approach to determine, approximately, the correlation time for a given variable. The idea of this analysis is to plot:

$$\sigma_{\bar{A}}^2 \text{ as a function of } \frac{\sigma_A^2}{t_{\text{tot}}}$$

for simulations of different lengths t_{tot}. The slope of this line gives twice the correlation time, per the equation

$$\sigma_{\bar{A}}^2 = 2\tau_A \frac{\sigma_A^2}{t_{\text{tot}}}$$

In practice, we take a long simulation trajectory and first compute the following:

$$\sigma_A^2 = \text{variance of } A \text{ over entire simulation trajectory}$$

Then, we subdivide the trajectory into different, nonoverlapping time segments or **blocks.** We can then compute the other quantities above:

$$\bar{A}_i = \text{average } A \text{ for each block } i$$

$$\sigma_{\bar{A}}^2 = \text{variance of the } \bar{A}_i$$

$$t_{\text{tot}} = \text{length of each block } i$$

By performing the block averages for different numbers of blocks, and hence different t_{tot}, we are able to find the slope corresponding to τ_A above.

Multiple trials

While it is very important to perform averages for lengths that exceed correlation times in a single simulation, it is common practice to also perform multiple **trials** of the same run and average the results not only in time but also across the different trials. The use of multiple trials can help to

produce results that are more statistically independent. Each trial should be seeded with a different random initial velocity set.

Notation

In the remainder of these notes and in later lectures, we will drop the notation $\bar{A}$ and use $\langle A \rangle$ to designate both true equilibrium, statistical-mechanical averages *and* finite-duration simulation averages. Keep in mind, though, that any average computed from simulation will be subject to the statistical properties described above.

Transport properties

As NVE molecular dynamics simulations follow the Newtonian evolution of the atomic positions, they give rise to trajectories that accurately represent the true dynamics of the system. Thus, these simulations can be used to compute kinetic transport coefficients in addition to thermodynamic properties.

Self-diffusivity: Einstein formulation

The self-diffusion constant D is defined as the linear proportionality constant between the mass/molar flux of a species and the concentration gradient (Fick's law). For a uniform diffusion constant (with space, as in a homogeneous bulk phase), the following equation defines evolution of the concentration ρ (molecules per volume) with time:

$$\frac{\partial \rho(\mathbf{r},t)}{\partial t} = -D\nabla^2 \rho(\mathbf{r},t)$$

We can rewrite this equation in terms of the probability density that we will find a molecule at some point in space. Letting $\wp(\mathbf{r};t)$ be this probability, we then have

$$\rho(\mathbf{r},t) = \wp(\mathbf{r};t)N$$

$$\int \wp(\mathbf{r};t)d\mathbf{r} = 1$$

Making this substitution,

$$\frac{\partial \wp(\mathbf{r};t)}{\partial t} = -D\nabla^2 \wp(\mathbf{r};t)$$

Imagine that a molecule is known to initially start at a given point $\mathbf{r} = \mathbf{r}_0$ in space at $t = 0$. Then, the solution to $\wp(\mathbf{r};t)$ is given by

$$\wp(\mathbf{r};t) = (\pi D t)^{-\frac{3}{2}} \exp\left(-\frac{|\mathbf{r}-\mathbf{r}_0|^2}{4Dt}\right)$$

We can compute from this the **mean-squared displacement** with time:

$$\langle |\mathbf{r}-\mathbf{r}_0|^2 \rangle = \int \wp(\mathbf{r};t)|\boldsymbol{r}-\boldsymbol{r}_0|^2 d\mathbf{r}$$
$$= 6Dt$$

In other words, the mean-squared displacement grows linearly with time with a coefficient of $6D$. This equation is an **Einstein relation**, after Albert Einstein's seminal work in diffusion. Importantly, it gives us a way to measure the diffusion constant in simulation:

1. At time $t = 0$, record all particle positions $\mathbf{r}_0^N$.

2. At regular intervals t, compute the mean squared displacement averaged over all atoms, $|\mathbf{r}-\mathbf{r}_0|^2$.

3. Find the diffusion constant from the limit at large times:
$$D = \lim_{t\to\infty} \frac{\langle |\mathbf{r}-\mathbf{r}_0|^2 \rangle}{6t}$$
or, better, from the slope of the mean squared displacement at long times:
$$D = \frac{1}{6} \lim_{t\to\infty} \frac{d}{dt} \langle |\mathbf{r}-\mathbf{r}_0|^2 \rangle$$

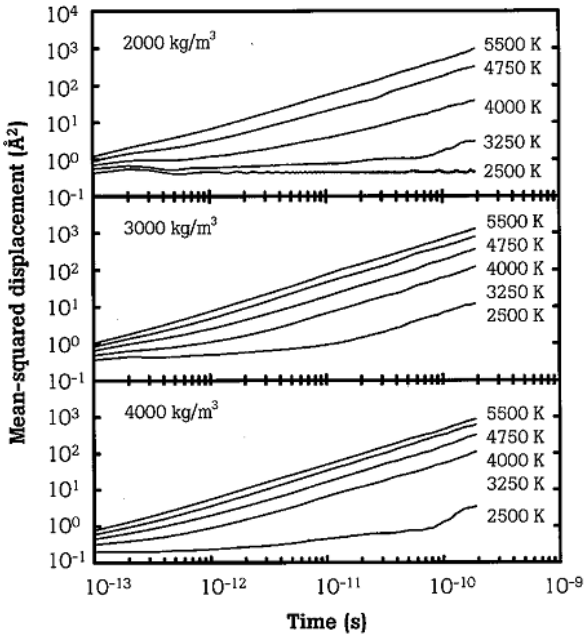
Some logistical aspects must be kept in mind:

- The time at which the diffusion coefficient is measured should be a number of relaxation times of the system.

- For better statistics in computing the mean-squared displacement curve (vs. time), it is often useful to have **multiple time origins,** e.g., $\mathbf{r}_0^N, \mathbf{r}_1^N, \mathbf{r}_2^N, \ldots$ reference positions taken at statistically independent time intervals (i.e., a relaxation time). Then, at each time t one can make updates to the average mean-squared displacement curve at times $t - t_0, t - t_1, t - t_2, \ldots$ using the respective reference coordinates.

- If a system consists of multiple atom types, each can have its own self diffusion coefficient and the equations will involve separate mean squared displacement calculations for the respective atoms of each type.

The following shows the mean-squared displacement curves for oxygen atoms in liquid silica (SiO_2), taken from [Shell, Debenedetti, Panagiotopoulos, Phys. Rev. E 66, 011202 (2002)]:



Notice the log-log plot. We see linear behavior in the curves (expected for random-walk diffusion according to a diffusion constant) after some initial time period has passed. There are different regimes in particle diffusion:

- **ballistic regime** – At very short times, particles do not "feel" each other, $\mathbf{r} \approx \mathbf{v}t$, and the mean squared displacement simply scales as $|\Delta\mathbf{r}|^2 \sim v^2 t^2$. On the plot above, we would expect to see a slope of 2 at short times, $\ln|\Delta\mathbf{r}|^2 \sim 2\ln t$.

- **diffusive regime** – At long times, particles have lost memory of their initial positions and are performing a random walk according to the diffusion constant, $|\Delta\mathbf{r}|^2 \sim 6Dt$. We should only use data from this regime when computing the diffusion constant. Notably, the slope in this regime on the above plot should be 1, $\ln|\Delta\mathbf{r}|^2 \sim \ln t$.

- **caged regime** – At intermediate times, the mean squared displacement may not follow either of these scaling laws. Often, $|\Delta\mathbf{r}|^2$ will appear to plateau for some time period. This behavior is typical of sluggish dynamics in viscous liquids and polymers.

Self-diffusivity: Green-Kubo formulation

It is entirely possible to transform the Einstein expression for the self-diffusivity, in terms of the mean squared displacement, into a form that relates to the atomic velocities instead, using

$$|\mathbf{r} - \mathbf{r_0}|^2 = \left| \int_0^t \mathbf{v}(t) dt \right|^2$$

We substitute this expression into the equations above and simplify using ideas similar to those developed in the time-correlation section. This approach gives a **Green-Kubo** relation that connects the diffusivity to the **velocity autocorrelation function**:

$$C_{\mathbf{v}}(t) = \frac{\langle \mathbf{v}(t) \cdot \mathbf{v}(0) \rangle - \langle \mathbf{v}(0) \rangle \cdot \langle \mathbf{v}(0) \rangle}{\langle \mathbf{v}(0) \cdot \mathbf{v}(0) \rangle - \langle \mathbf{v}(0) \rangle \cdot \langle \mathbf{v}(0) \rangle} = \frac{\langle \mathbf{v}(t) \cdot \mathbf{v}(0) \rangle - \langle \mathbf{v}(0) \rangle \cdot \langle \mathbf{v}(0) \rangle}{\sigma_{\mathbf{v}}^2}$$

The averages here are performed for particles of the same type and over multiple time origins for recording the initial velocity $\mathbf{v}(0)$. The diffusion constant relates to the integral of $C_{\mathbf{v}}$:

$$D = \frac{\sigma_{\mathbf{v}}^2}{3} \int_0^\infty C_{\mathbf{v}}(t) dt$$
$$= \frac{\sigma_{\mathbf{v}}^2 \tau_{\mathbf{v}}}{3}$$

In other words, the diffusion constant relates to the correlation time of the velocity.

In practice, the autocorrelation function is approximated by a discretized array (the index corresponding to a time bin) and computed in a similar manner as the mean-squared displacement using multiple time origins. This function typically decays to near zero in a finite length of time and thus the integral only needs to be computed up until this point. Sometimes special techniques are needed to coarse-grain time in order to treat the statistical fluctuations around zero in the tails of the computed autocorrelation function.

Other transport coefficients

A very general theory shows that Green-Kubo relations can be formulated for any transport coefficient that is a linear constant of proportionality between a flux and a gradient. Some examples include the **bulk viscosity**, **shear viscosity**, the **thermal conductivity**, and the **electrical conductivity.** Expressions for these can be found in standard texts. The bulk viscosity, for example, is given by:

$$\eta_V = \frac{\sigma_{PV}^2}{V k_B T} \int_0^\infty C_{PV}(t) dt$$

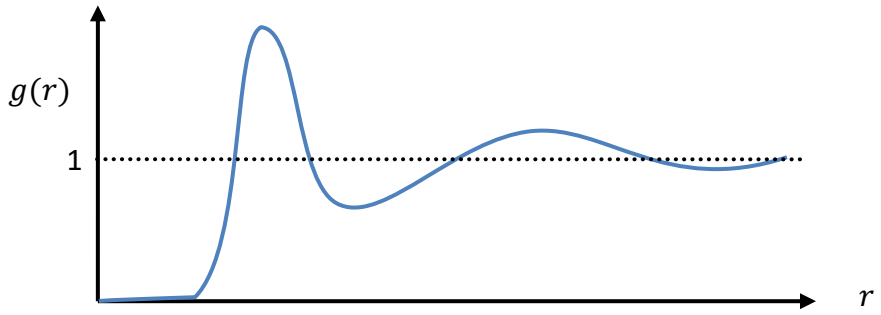
where $C_{PV}(t)$ is the correlation function for fluctuations in the term PV:

$$C_{PV}(t) = \frac{\langle PV(t) \cdot PV(0) \rangle - \langle PV(0) \rangle \cdot \langle PV(0) \rangle}{\langle PV(0) \cdot PV(0) \rangle - \langle PV(0) \rangle \cdot \langle PV(0) \rangle} = \frac{\langle PV(t) \cdot PV(0) \rangle - \langle PV(0) \rangle \cdot \langle PV(0) \rangle}{\sigma_{PV}^2}$$

Structure-based averages

Radial distribution functions (RDFs)

The **radial distribution function (RDF)** or **pair correlation function** is a measure of the structure of a homogeneous phase, such as a liquid, gas, or crystal. Given that a particle sits at the origin, it gives the density of particles at a radial distance r from it, relative to the bulk density.



Formally, the pair correlation function for a monoatomic system in the canonical ensemble is defined by:

$$g(\mathbf{r}_1, \mathbf{r}_2) = \frac{V^2(N-1)}{N} \frac{\int e^{-\beta U(\mathbf{r}^N)} d\mathbf{r}_3 d\mathbf{r}_4 \ldots d\mathbf{r}_N}{Z(T,V,N)}$$

where $Z(T,V,N)$ is the canonical partition function. In an isotropic medium, this function depends only on the relative distance between two atoms, not their absolute position:

$$g(r_{12})$$

For an ideal gas with $U(\mathbf{r}^N) = 0$,

$$g(r_{12}) = \frac{N-1}{N}$$
$$\approx 1 \quad (\text{large } N)$$

Note that,

$$\int (4\pi r^2 dr) \rho g(r) = N - 1$$

One can also define a radial distribution function for atoms of different types, e.g., between hydrogen and oxygen atoms in liquid water. In this case, we can define

$$g_{AB}(\mathbf{r}_1, \mathbf{r}_2) = V^2 \frac{\int e^{-\beta U(\mathbf{r}^N)} d\mathbf{r}_3 d\mathbf{r}_4 \ldots d\mathbf{r}_N}{Z(T, V, N_A, N_B)}$$

for two atom types A and B.

RDFs can be computed using **histograms** of the pairwise distances between particles. For a monatomic system with just one kind of particle, the recipe is the following:

1. At periodic intervals in the simulation, examine all pairwise $N(N-1)/2$ distances of the N particles. One does not need to examine every time step, but only those approximately spaced by the relaxation time in the system, or a moderate fraction thereof. Let the number of these intervals be n_{obs}.

2. Let c_i denote an array of histogram counts for the total number of times a pairwise distance r_{ij} is observed, where $i\delta \leq r_{ij} < (i+1)\delta$ and δ is the width of the histogram bins.

3. After sufficient data collection, the RDF can be approximated at discrete intervals $i\delta$. For atoms of the same type:

$$g_{AA}(i\delta) = \frac{c_i}{n_{\text{obs}} N_A (N_A - 1)/2} \times \frac{V}{\frac{4\pi\delta^3}{3}((i+1)^3 - i^3)}$$

For atoms of different types:

$$g_{AB}(i\delta) = \frac{c_i}{n_{\text{obs}} N_A N_B} \times \frac{V}{\frac{4\pi\delta^3}{3}((i+1)^3 - i^3)}$$

Energy and pressure from RDFs

For pair potentials, integrals of an RDF can be used to compute the potential energy and pressure:

$$\langle U \rangle = \frac{N}{2} \int_0^\infty [4\pi r^2 \rho g(r)] u(r) dr$$
$$= \frac{2\pi N^2}{V} \int_0^\infty r^2 g(r) u(r) dr$$

$$\langle P \rangle = \frac{N k_B T}{V} - \frac{N}{6V} \int_0^\infty [4\pi r^2 \rho g(r)] r \frac{du(r)}{dr} dr$$
$$= \frac{N k_B T}{V} - \frac{2\pi N^2}{3V^2} \int_0^\infty r^3 g(r) \frac{du(r)}{dr} dr$$

The latter equation is merely an extension of the virial expression for the pressure. If there are multiple atom types in the system, then we will have multiple $g(r)$ functions that need to be integrated. For example, for two types A and B:

$$\langle U \rangle = \langle U_{AA} \rangle + \langle U_{BB} \rangle + \langle U_{AB} \rangle$$
$$= \frac{2\pi}{V} \int_0^\infty r^2 [N_A^2 g_{AA}(r) u_{AA}(r) + N_B^2 g_{BB}(r) u_{BB}(r) + 2N_A N_B g_{AB}(r) u_{AB}(r)] dr$$

The coefficient of two in front of the AB terms comes from the fact that these interactions are *not* double counted when performing the usual integral. A convenient way to express this is through a double sum over all atom types (with M total types):

$$\langle U \rangle = \frac{2\pi}{V} \int_0^\infty r^2 \left[\sum_{X=1}^{M} \sum_{Y=1}^{M} N_X N_Y g_{XY}(r) u_{XY}(r) \right] dr$$

A similar expression can be derived for the pressure:

$$\langle P \rangle = \frac{k_B T N_{\text{tot}}}{V} - \frac{2\pi}{3V} \int_0^\infty r^3 \left[\sum_{X=1}^{M} \sum_{Y=1}^{M} N_X N_Y g_{XY}(r) \frac{du_{XY}(r)}{dr} \right] dr$$