# Lecture 2: Molecular Simulations

## I. Introduction

Molecular states are frequently described by statistical probabilities of ensemble averages. The theories of statistical mechanics or statistical thermodynamics give us expressions to describe a *whole population* of molecules and, in some cases, give us a hint at what individual molecules may be doing. But to investigate the conformational exploration of *single molecules* computer simulation is currently the best approach.

In the computer lab we will start with one of the simplest of molecular simulations: single-atom molecules with a single type of interaction potential. The basics of molecular simulation will be introduced with this system. Then we will move up in complexity to simulate a peptide fragment *in vacuo* with additional interaction potentials. Finally, we will simulate a protein in water solution with a "complete" set of empirical interaction potentials. Along the way we will use both Monte Carlo simulations and Molecular Dynamics simulations.

Statistical thermodynamics provides the theoretical basis for molecular simulations and appropriate (and very limited) *theoretical* topics in statistical thermodynamics will be covered in following lectures. In the computer lab usually we will be more concerned with the *practical* information needed to carry out and understand molecular simulations. The topics in this lecture fall into both categories. Inevitably there is a glossary of terms and concepts that one must know to proceed.

## II. Molecular Mechanics

**The fundamental procedures in molecular simulations are to move a collection of atoms repetitively in space and calculate their energy after each step.** Quantum mechanics methods would be the most rigorous way to calculate the energy of a collection of atoms but would take too long to be of practical value for molecular simulations given the current state of computer processing speed. Therefore, **molecular mechanics** was developed as a method to describe the energies of interacting atoms using more rapid, but less accurate, empirical procedures. A **molecular mechanics model** typically consists of:

- Atoms as spheres
- Bonds between the atoms as springs
- Atomic interaction energies as empirically derived analytical functions

Since our initial simulations will be of single-atom molecules we will delay a discussion of bonds and their energies. Also, our initial simulation molecules will be neutral in charge so we will ignore electrostatics due to permanent charge-charge interactions for now. For our first molecular mechanical model we will use only two types of interaction energies; one attractive and one repulsive. These potential energies will constitute the entire **force field** for the first simulations. Later in the course we will become familiar

with more complete empirical force fields that contain parameters for atomic bonds, angles, torsions, and charge-charge interactions as well as the attractive and repulsive interactions we consider now.

**1. van der Waals Attractions**

A full derivation of the expressions that describe van der Waals interactions would require a very large examination of the history of physical chemistry. *In this section the points to be made are simply that all atoms have fluctuating partial electronic charges; these cause favorable interatomic interaction; and the interaction falls off over very short distances*.

The term "Van der Waals interactions" should refer to all non-covalent interactions between molecules. In computational biology we usually make a distinction between interactions that involve *permanently* charged species and call these **electrostatic**, with **van der Waals** referring only to interaction of *fluctuating* partial charges. But this duality is an arbitrary distinction.

We are familiar with the fact that two charged molecules may be repulsed by, or attracted to, each other depending on whether they have the same or opposite charges associated with them. The potential energy of interaction between fully charged species is most simply described by a **Coulombic** expression

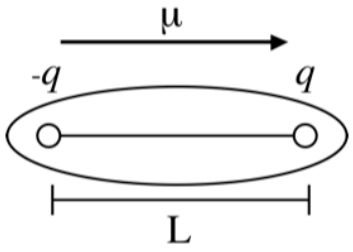
$$E_r = q_1 q_2 / 4\pi\varepsilon_r\varepsilon_0 r \qquad (1)$$

where $q$ is the charge on each species, $\varepsilon_r$ is the relative permitivity of the medium separating the charges, $\varepsilon_0$ is the permittivity in a vacuum ($8.85 \times 10^{-12}$ $C^2/Jm$), and $r$ is the distance between the charges. The relative permitivity may be thought of as the polarization of the medium, i.e. a measure of how well it can screen the interaction between the charges compared to a vacuum. It is sometimes called the **dielectric constant** of the medium.

Even neutral molecules usually have fluctuating internal partial charge separation. Molecules with internal partial charges inherently have a **dipole moment** that reflects the vectorial separation of partial charges within the molecule. A dipole moment is described by

$$\mu = qL \qquad (2)$$

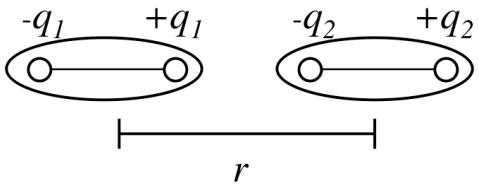
where $L$ is the internal separation between partial charges and $q$ is the absolute value of the partial charge.



Two molecules with dipole moments may be repulsive or attractive to each other depending on the relative orientations of their dipole moments. If the positive ends are closer, they repel and if one positive end is closer to the negative end of the other molecule they attract. The potential energy between two molecules with the same dipole moment in collinear arrangement of the dipoles is

$$E_r = (\mu_1\mu_2/4\pi\varepsilon_r\varepsilon_0 r^3) \qquad (3)$$

This expression only applies to polar molecules in collinear orientation as shown below but it makes the general point.




The general point here is that the distance dependence in equation (3) is $\mathbf{1/r^3}$ and the distance dependence in equation (1) is $\mathbf{1/r}$. The interaction between dipoles falls off faster than between point charges because the molecules "see" both the + and - partial charges and they begin to appear neutral to each other at large separations.

The above interaction is due to the electric fields of the two dipoles although we model the interaction using simple point charges. The field generated by a dipole moment is the sum of the field generated by each partial charge. An expression for the magnitude of a dipole electric field at some distance **r** is

$$Field_r = 2\mu/4\pi\varepsilon_r\varepsilon_0 r^3 \qquad (4)$$

Again, this expression is for the field of dipole $\mu_1$ as seen by dipole $\mu_2$ above in collinear arrangement. Notice the $\mathbf{1/r^3}$ dependence: The electric field also falls off with the cube of the distance.

The electronic configuration of even nonpolar molecules is constantly fluctuating and therefore transient dipole moments are created. If the molecules were freely rotating, then the interaction between dipoles would average to zero. However, these transient dipole moments create transient electric fields and the fields affect the molecules nearby as well as causing attraction/repulsion. In fact, the transient fields of two neighboring molecules will tend to orient the transient dipole moment vectors in the two molecules into the collinear arrangement illustrated above. This combination of effects results in a lower, more favorable energy of interaction between the dipoles.

These induced-dipole to induced-dipole interactions are variously called **dispersion** or **London** interactions. *They result in the fact that all molecules, even neutral molecules, have an attractive force between them*. Sometimes this universal attractive force is overshadowed by other interactions (such as charge-charge repulsion) but this universal attractive force is important for molecular interaction and results, for example, in the

condensed phases of nonpolar substances such as benzene. Today the term, van der Waals interactions, commonly refers to these dispersion interactions.

The dispersion interactions are doubly dependent on distance between the dipoles described above. Both the orienting effect of the field and the potential energy of interaction fall off with their respective $\mathbf{1/r^3}$ dependencies. *The total dispersion effect falls off with a distance dependence of the product of both dependencies,* $\mathbf{1/r^6}$. Therefore, these interactions are considered short-range interactions.

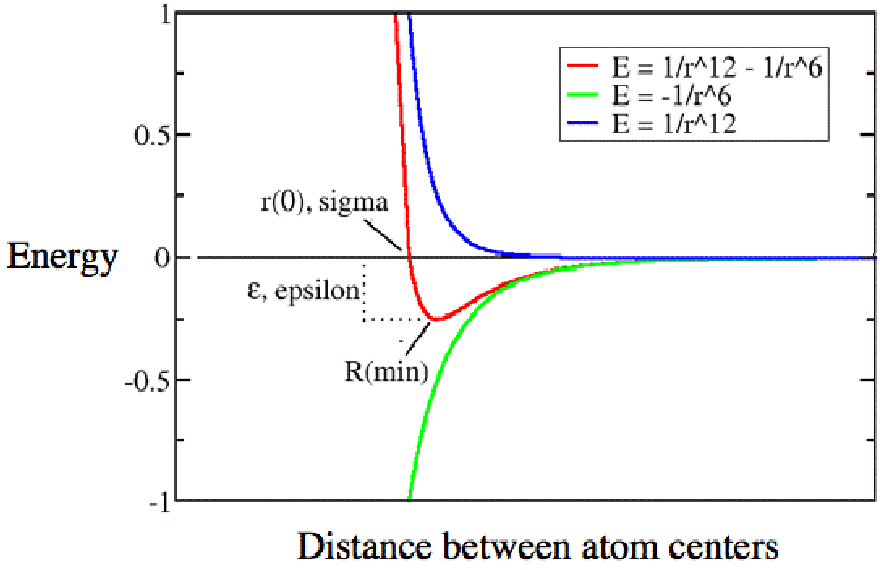
**2. Lennard-Jones Potential**

When atoms are forced into close proximity the electronic interactions give rise to an increasing repulsion. An exponentially rising repulsive energy empirically describes this property well in most cases. A common practice is to use a $\mathbf{1/r^{12}}$ function where $\boldsymbol{r}$ is the distance between atoms centers. The Lennard-Jones interaction potential combines this *repulsive* $\mathbf{1/r^{12}}$ function with a $\mathbf{-1/r^6}$ *attractive* function to capture both the repulsive energy and the above described attractive van der Waals energy in one function.

$$E_r = 4\varepsilon\{(r_0/r)^{12} - (r_0/r)^6\} \qquad (5)$$

where $r_0$ (sometimes called sigma, $\sigma$) is the distance at which the repulsive and attractive forces are exactly balanced and the interaction energy equals zero, and epsilon ($\varepsilon$) is the magnitude of the energy minimum (described below). The figure below shows the repulsive function, the attractive function and their *combination* to form a **Lennard-Jones potential function** between two identical atoms.

Notice that the repulsive potential falls off faster than the attractive potential. This means that at some intermediate separation the attractive potential is dominant and there is a favorable minimum in the energy.



The distance at the energy minimum of the Lennard-Jones potential is the *equilibrium distance between the two atoms* and one half this distance is often called the **van der Waals radius** of the atoms ($R_{min} = 2^{1/6} r_0 = 1.12 r_0$ ). One half the distance of $r_0$ (sigma, $\sigma$) is often called the **hard shell radius** of the atom.

To implement the Lennard-Jones potential in a molecular mechanics model one assigns empirically derived values for $\varepsilon$ and $r_0$ that best describe the interaction energy between an atom type and itself. For example, an accepted value for $\varepsilon$ between two Argon atoms is 0.238 kcal/mol and the corresponding $r_0$ is 3.4 Å. The first molecular simulation using empirical parameters such as these for the Lennard-Jones interaction potential was by A. Rahman in 1964. (This paper is available in PDF format in the References section on the course web site). Later when we consider interaction energies between two different atoms we will learn special rules to obtain the average $\varepsilon$ and $r_0$ for applying the Lennard-Jones potential energy function to these cases. Note that in some texts $r_0$ is called $\sigma$.

(Application Note: In the Lennard-Jones potential, atoms that are far from the atom of interest have vanishingly small energies of interaction. To speed up a simulation it is standard practice to cut off the interatomic interaction calculation at some distance (8 – 12 Å). This saves a great deal of time because the number of pairwise interactions increases as the third power with distance. A small correction is added to the energy calculation to account for the atoms outside the cutoff distance.)

The choosing of values for $\varepsilon$ and $r_0$ is both an art and an empirical science. The bottom line is that one would like to choose values that give rise to simulations which mimic nature. But it has turned out that the best choice of force field potential parameters to mimic one aspect of nature, liquid density say, frequently is not the best choice to mimic something like structure or conformational fluctuation. There is still room for much research in this area.

Finally, it is worth pointing out that the LJ potential estimates a potential energy *not* a free energy. We will come back to this distinction later in the course.

**III. Monte Carlo Simulations**

The idea of Monte Carlo (MC) simulations is to randomly sample all possible microstates of the system (a collection of atoms), and to collect appropriate states so that one obtains a population of states that is distributed according to a **Boltzmann distribution**.

$$P(state) = \frac{e^{-\beta E(state)}}{Z} \qquad (6)$$

where $P(state)$ is the probability of finding a specific system state, $E(state)$ is the energy of the state, $\beta = 1/kT$ ($k$ = the Boltzmann constant and $T$ is the temperature in Kelvin) and $Z$ is the partition function (described later).

We will spend much time talking about the Boltzmann distribution because this concept is the basis for statistical thermodynamics and much of molecular simulation. For now

suffice it to say that this distribution is such that lower energy (more favorable) states make up most of the population at room temperature and that the relative abundance of states is inversely proportional to their respective energies. A Boltzmann distribution, by definition, is at equilibrium.

The sequence of events in a Monte Carlo simulation move is (given in *pseudocode*):

1. Calculate the initial (old) **potential** energy of the system ($E$)
2. Randomly move one or more molecule(s) in the system
3. Calculate the change in energy of the system, $\Delta E$
4. if $\Delta E < 0$:
   accept the move; new energy becomes old
   go to step 2
5. else compute a random number *Ran* between 0 and 1 and then ask
   if *Ran* < $e^{-\Delta E/kT}$
      accept the move; new energy becomes old
   else
      reject the move and keep the original system state
   go to step 2

Steps 4 and 5 represent application of the **Metropolis criterion**. Step 5 allows for acceptance of higher energy, unfavorable moves. Thus the system does not just go down an energy gradient to the lowest possible energy (as in pure energy minimization schemes), but rather, it explores all possible energy states.

Repetitively performing these steps accomplishes the generation of a population distributed according to the relative energies as expressed in the Boltzmann probability expression given above. But the algorithm also accomplishes this with an important condition: **detailed balance**.

**Detailed Balance**. Obviously, if the ensemble distribution is at equilibrium we do not want the accepted moves to destroy the equilibrium. This means that the average number of accepted moves that leave state ***i*** must equal the average number of accepted moves that enter state ***i*** from all other states ***j…n***. *This is the detailed balance condition*. Another way to state it is that the number of systems in ***i*** relative to the number of systems in ***j*** remains constant. Therefore

$$P_i \ [P_{trans}(i \rightarrow j)] = P_j \ [P_{trans}(j \rightarrow i)] \qquad (7)$$

where $P_i$ is the probability of state ***i*** existing and $P_{trans}(i \rightarrow i)$ is the transition probability of state ***i*** becoming state ***j***. Rearranging equation (7) leads to,

$$P_i / P_j = [P_{trans}(j \rightarrow i)] / [P_{trans}\text{->}(i \rightarrow j)] \qquad (8)$$

In a very famous paper a famous group of people (Metropolis, N. Rosenbluth, A.W., Rosenbluth, N.N., Teller, A.N. and Teller, E. Equation of state calculations by fast computing machines. J. Chem. Phys.

21:1087-1092, 1953) showed that the steps outlined in the algorithm above (the Metropolis criterion) satisfy this condition. Substituting from equation (6)

$$\frac{P_i}{P_j} = \frac{e^{-\beta E_i}}{e^{-\beta E_j}} = e^{-\beta(E_i - E_j)} \qquad (9)$$

The relative probabilities of states ***i*** and ***j*** are proportional to the exponential of their energy differences. This is another way of stating that the probabilities follow Boltzmann's law. *If detailed balance were not obeyed a Boltzmann distribution would not necessarily be obtained*.

**IV. Molecular Dynamics Simulations**

The idea of molecular dynamics (MD) simulations is to compute the equilibrium position and transport properties of a system over time by solving Newton's equations of motion. The trajectory of microstates of the system over time represents a population of states that is distributed according to a Boltzmann distribution. Molecular dynamics simulations provide information on both the **positions** and **momenta** of the atoms in the system in contrast to MC where only the **positions** are relevant. The basic steps in an MD simulation are to start with a collection of atoms for which we have assigned individual random velocities (of appropriate average magnitude, see below) and then,

1. Move all atoms very small time step according to their respective velocities
2. Calculate the new forces on each atom due to interaction with all other atoms
3. Calculate changes in velocities due to new forces using Newton's second law
4. Update positions as in step 1

Temperature and kinetic energy play important roles in molecular dynamics so we start the glossary of terms and concepts relevant to MD with these two quantities.

The **equipartition theorem** tells us that each translational degree of freedom ($x$, $y$, $z$) of an ideal gas atom has the same average energy $\langle \mathbf{1/2}\ m\mathbf{v}^2 \rangle$ where ***m*** = mass and ***v*** = velocity and that the overall mean kinetic energy of an atom is **3/2 *kT***, where ***k*** = the Boltzmann constant and ***T*** = temperature. For each of the three translational degrees of freedom ($x$, $y$, $z$)

$$\langle \mathbf{1/2}\ m\mathbf{v}^2 \rangle = \mathbf{1/2}\ kT \qquad (10)$$

and therefore the total kinetic energy of a system of N atoms defines the instantaneous temperature

$$T = \sum_{i=1}^{N} \left[ m_i v_i^2 / kN \right] \qquad (11)$$

The point here is: *In molecular dynamics calculations the temperature is a measure of atomic velocities*. When we assign the initial velocities to each atom at the beginning of the simulation we choose the random distribution of velocities so that the average is

equivalent to the requested temperature.

Application Note: There exist several algorithms to update the positions and momenta of atoms in MD. Below is one example – you are not responsible for this on the exams; it is included as an extra for those who are interested.

At each step for each pair of atoms close enough to interact we compute the force between them. For example

$$f(r) = -\partial E(r)/\partial r \qquad (12)$$

which for an atomic Lennard-Jones interaction force, given equation (5), may be differentiated to

$$= 48\varepsilon/r^2 \left[(r_0/r)^{12} - 0.5\,(r_0/r)^6\right] \qquad (13)$$

Newton's second law is

$$f(r) = ma \qquad (14)$$

rearranging

$$a = f(r)/\,m$$

where

$$a = d^2r/dt^2 = \text{acceleration}$$

With the forces in hand we integrate these equations of motion and update the positions for a time step $\Delta t$. The new velocity after the time step is

$$v(t+\Delta t) = v(t) + a\Delta t = v(t) + d^2r/dt^2\,\Delta t \qquad (15)$$

The new position is then

$$r(t+\Delta t) = r(t) + v(t+\Delta t)\,\Delta t \qquad (16)$$

The above procedure is conceptually the most straightforward. However integration of Newton's equations may be done with a variety of other algorithms with names such as Verlet, Leap-Frog, Gear-Predictor. These have different advantages and are described in the references.

There is one point about the MD updating of atom positions that you should remember. In order for the above to work without the velocities getting out of control, and to sample the fastest molecular motions, the time step must be very small. Usually 1 fs ($10^{-15}$ sec) is used and in this time an atom only moves about *1/500 of its diameter*. Thus to explore the total phase space of positions and momenta *very* long simulations are necessary.

## V. Equilibrium, Conformational Search and Randomness

In addition to detailed balance in MC simulations (Section III, above) an important concept in *both* MC and MD molecular simulations is **ergodicity**. The **ergodic hypothesis** states that every accessible state of the system should be reached in a finite number of simulation steps from any other accessible state. This means the algorithm is ergodic. In other words, over a very long period of time, a single particle is expected to sample all the states that would be found in an instantaneous sample of a large (e.g. $N_A$) number of particles. The ergodic hypothesis further implies that to compute the average

of a function of the coordinates of a many-particle system we can either compute that quantity by averaging over time the evolution of a single system or by averaging over different systems of an ensemble. In Monte Carlo (MC) we are averaging over different systems, in molecular dynamics (MD) we are averaging a single system over time.

One other condition that is assumed to be true in all molecular simulations is that after a sufficient amount of time (MD) or number of moves (MC) the average of an ensemble function is independent of the initial conditions. In other words it is assumed that the move set represents a **Markov chain** or **random walk**.

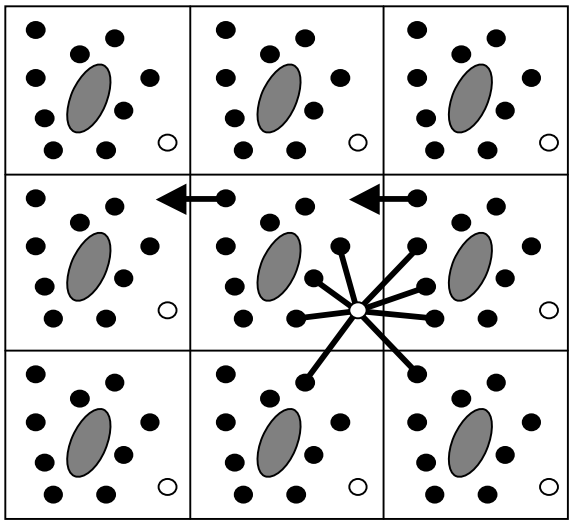
To restate: Two assumptions that are important in *both* MC and MD are ergodicity and that the simulations represent a Markov chain. Detailed balance applies only to MC.

Important questions one must ask are: Have we carried out the simulation for a sufficient number of steps so that the ensemble is independent of the initial state? Have we sampled enough states to satisfy the ergodic hypothesis? *In other words have we reached equilibrium?* This condition must be true if the simulation is a Markov process and the ergodic hypothesis applies. There are sophisticated methods to determine if equilibrium has been reached and there are some common sense tests one can apply. We will explore this question in the computer lab. *The question of equilibrium is important because if the system is not at equilibrium the ensemble will not be Boltzmann distributed* (in which case we can consider it garbage).

**VI. Periodic Boundary Conditions**

We have talked about ensembles of systems (where a **system** is a group of atoms) but we have not yet mentioned the appropriate number of atoms in a system for simulation. To avoid **edge effects** one would like to simulate a portion of an infinite system. This may be achieved by using **periodic boundary conditions**. A volume containing hundreds or thousands (or recently, millions) of atoms is taken to be the primitive of an infinite periodic lattice of identical volumes. We will use cubic volumes in our computer experiments but the method is not limited to a cubic lattice. The figure below shows a periodic boundary system in two dimensions. The system contains large elliptical and small spherical atoms. The central square represents our primitive volume. Notice that the open circle atom interacts with other atoms in its own square and with atoms in adjoining squares - there are no effective edges. Furthermore as one atom diffuses out of the central square (say towards the left) it enters the adjoining square on the left and an equivalent atom enters the central square from adjoining square on the right. Since the atomic interactions we care about are over relatively short distances we can use periodic boundary conditions to mimic an infinite system. Exactly how this is done will be described in the first programs used in the lab.



**VI. Various Simulation Ensembles**

The word **ensemble** refers to all the parts taken together. We have used ensemble to mean a collection of groups of atoms in different states. But for molecular simulations the word ensemble takes on special meaning.

Consider four different thermodynamic environments:

(a) a closed, isolated isoenergetic system ($N$, $V$, $E$ constant, where $N$ = number of particles, $V$ = volume, $E$ = energy)
(b) a closed, isothermal system ($N$, $V$, $T$ constant, where $T$ = temperature)
(c) an open, isothermal system ($\mu$, $V$, $T$ constant, where $\mu$ = chemical potential)
(d) a closed, isobaric isothermal system ($N$, $P$, $T$ constant, where $P$ = pressure)

These thermodynamic environments are usually called the
**microcanonical ensemble** (*NVE*),
**canonical ensemble** (*NVT*),
**grand canonical ensemble** ($\mu VT$) and
**isobaric ensemble** (*NPT*)

In the discussion above on MC we assumed that the temperature ($T$) remained constant and for MD we assumed that the energy ($E$) remained constant. In both the number of atoms ($N$) and the volume ($V$) remained constant. One may say that the MC and MD simulations were carried out in two different ensembles; the *NVT* or *canonical* ensemble for MC and the *NVE* or *microcanonical* ensemble for MD. Most biological systems are simulated using the isobaric ensemble at constant pressure (*NPT*).

**VII. When to use MC and when to use MD**

*Both MC and MD create a Boltzmann distribution of states of the system*. (This is the first purpose of performing molecular simulations). The choice of which method to use for typical simulations of biological molecules is largely one of past experience and available

tools. MC requires less computation per cycle because only the potential energy and not velocities have to be calculated. But because MC chooses steps in a somewhat inefficient random way the number of cycles required to obtain uncorrelated samples may be greater than with MD. Both MC and MD may be carried out in the isobaric ensemble (*NPT*) that is the one most akin to biological experiments. For special purposes; for example to calculate time dependent quantities or to use a specific ensemble, there may be other considerations that bear on the choice of methods and these are discussed in the reference literature.

**VIII. Reduced Units.**

If we were to use real (SI) units for the parameters of molecular interactions and time scales we would have typical interaction distances of $10^{-10}$ m, energies of interaction per molecule of $10^{-20}$ cal, and time steps of $10^{-15}$ sec. Instead we will choose convenient units of energy, length and mass on a **molecular scale** and derive all other quantities in terms of these convenient units. This molecular scale is in terms of **reduced units**. The basic molecular scale units for argon (Ar) are defined as follows (where $r$ = measured SI distance; $E$ = calculated SI energy):

- unit of length, $r^*$ $= r/\sigma$ ($1\sigma = 3.4$ Å, $3.4 \times 10^{-10}$ m)
- unit of energy, $E^*$ $= E/\varepsilon$ ($1\varepsilon = 238$ cal/mol, or $k \cdot 120$°K, where $k = 1.9872$ cal/mol/K)
- unit of mass, $m^*$ $= m/M$ ($1m = 0.03994$ kg/mol)

With reduced units we can use a simplified form of the Lennard Jones potential. We define the reduced pairwise interaction potential, $E^* = E/\varepsilon$, and the reduced distance, $r^* = r/\sigma$. Note that the * indicates reduced units. Then the LJ potential has the form

$$E^* = 4[(1/r^*)^{12} - (1/r^*)^6]$$

The form of the LJ potential you will see in the code for your simulations uses reduced units. Other quantities used in simulations can also be stated in terms of σ, ε, $m$ (note that the * indicates reduced units). For Argon simulations the following are reduced units:

- $t^*$; unit of time $= t/\text{sqrt}[\varepsilon/m\sigma^2]$ ($1.099 \times 10^{-14}$ sec)
- $T^*$; unit of temperature $= Tk/\varepsilon$ (119.8 K)
- $P^*$; unit of pressure $= P\sigma^3/\varepsilon$ (41.9 MPa)
- $\rho^*$; unit of density $= \rho\sigma^3$ ($1680$ kg/m$^3$) (Note that this is *mass density*; we will also use the concept of volume density in setting up our simulations).

To compare our results from simulations with reduced units to experimental values we can use the above conversions.

**Keywords.** You should be able to define or explain each of these terms with a short phrase, mathematical expression or sentence.

Boltzmann distribution
canonical ensemble
Coulomb expression
Detailed Balance
dielectric constant
dipole moment
dispersion interactions
equipartition theorem
ergodic hypothesis
force field
grand canonical ensemble
hard shell radius
isobaric ensemble
Lennard-Jones Potential
London interactions
Markov chain
Metropolis criterion
microcanonical ensemble
Molecular Dynamics Simulation algorithm
molecular mechanics
molecular mechanics model
Monte Carlo Simulation algorithm
periodic boundary conditions
positions versus momenta
random walk
reduced units
van der Waals Attractions
van der Waals radius