THESIS TITLE

SEBASTIAN SALASSI

Supervisor: Dr. Giulia Rossi

Advisor: Prof. Annalisa Relini

What do you get if you multiply six by nine?
Six by nine. Forty two.
That's it. That's all there is.
I always thought something was fundamentally wrong with the universe.

Douglas Adams – The Restaurant at the End of the Universe

INTRODUCTION

LIST OF ACRONYMS

CG Coarse–Grained

COM Center of Mass

CV Collective Variable

DMPC 1,2–dimyristoyl–*sn*–glycero–3–phosphocholine

DNA Deoxyribonucleic Acid

DOF Degrees of Freedom

ESM Ewald Summation Method

FES Free Energy Surface

FF Force Field

FFT Fast Fourier Transform

MD Molecular Dynamics

PBC Periodic Boundary Conditions

PEF Potential Energy Function

PME Particle Mesh Ewald

PW Polarizable Water

RMS Root–Mean–Square

WHAM Weighted Histogram Analysis Method

CONTENTS

1 INTRODUCTION TO MOLECULAR DYNAMICS

For macroscopic bodies, the motion of a system in time and space is governed by the classical equations of motion, say Newton's laws, while reducing time and space scales, quantum mechanics kicks in. Despite the latter statement, classical laws of motion have proved to be a good approximation also at the molecular level, as long as atoms are massive enough.

In order to predict the time evolution of a complete system, such as the biomolecular system we will treat in this thesis, Newton's equations of motion need to be integrated numerically. The necessity of a numerical integration arises from the complexity of the interactions involved in realistic systems, often nonlinear functions of positions and momenta of the particles, which makes it impossible to obtain an analytical solution for the equations of motion.

In the first part of this Chapter the laws of classical and statistical mechanics will be briefly summarized. Then we will introduce the computational Molecular Dynamics (MD) method and analyze the main aspects of this technique with some details about the *empirical* Force Field (FF): the container of system model and parameters under study. This includes the way to consider and treat the inter–particle interactions at MD level. Then, coarse–graining procedure are introduced, with particular attention to the MARTINI FF, the main FF used in this thesis work. Lastly advanced sampling techniques are explained: the *Umbrella Sampling* and *Metadynamics*. This introductory Chapter are based on the books of Tuckerman [22], Leach [13], Frenkel and Smit [7] and Allen and Tildesley [1] to which the reader is addressed for a more complete discussion.

1.1 REVIEW OF CLASSICAL MECHANICS

Let us consider a system of N particles with mass m_i and coordinates $\vec{r}_1, \ldots, \vec{r}_N$. According to the Newton's second low each particle will experience a total force $\vec{F}_i$ such that

$$m_i \ddot{\vec{r}}_i = \vec{F}_i(\vec{r}_1, \ldots, \vec{r}_N) \tag{1.1.1}$$

The total force on each particle is defined as

$$\vec{F}_i(\vec{r}_1, \ldots, \vec{r}_N) = \vec{f}_i^{(e)}(\vec{r}_i) + \sum_{i \neq j}^{N} \vec{f}_{ij}^{(i)}(\vec{r}_i - \vec{r}_j)$$

where $\vec{f}_i^{(e)}$ is an external force acting on particle i and $\vec{f}_{ij}^{(i)}$, which in general depends only on distance between particle i and j, is the inter-particle force that i

exerts on j and *vice-versa*. Equations (1.1.1) are refereed to as *the equations of motion* of the system; integrating them, with the sets of the *initial conditions* at start time t_0 $\vec{r}_1(t_0), \ldots, \vec{r}_N(t_0)$ and $\dot{\vec{r}}_1(t_0), \ldots, \dot{\vec{r}}_N(t_0)$, the positions and the velocities of all the particles in the system at any time are known.

Another way to write the equations of motion is to use particles momenta $\vec{p}_i = m_i \dot{\vec{r}}_i$ and then the equations (1.1.1) become

$$\frac{d\vec{p}_i}{dt} = \vec{F}_i(\vec{r}_1, \ldots, \vec{r}_N) \tag{1.1.2}$$

The full set of 6N functions $(\vec{r}_1(t), \ldots, \vec{r}_N(t), \vec{p}_1(t), \ldots, \vec{p}_N)$ gives us a full description of the dynamics of the N–particle system. The set of functions above can be arranged in an ordered 6N–dimensional vector

$$\vec{x}(t) = (\vec{r}_1(t), \ldots, \vec{r}_N(t), \vec{p}_1(t), \ldots, \vec{p}_N) \tag{1.1.3}$$

called *phase space vector* or the *microstate* of the system at time t. All the possible microstates of a system generate a 6N–dimensional space called *phase space* of the system, indicated with Ω. Thus $\vec{x}(t)$ describes a particular trajectory in the phase space, i.e. the system evolution is the motion of a phase space point. For simplicity of notation let us introduce also the ordered vectors $\vec{r} = (\vec{r}_1, \ldots, \vec{r}_N)$, $\dot{\vec{r}} = (\dot{\vec{r}}_1, \ldots, \dot{\vec{r}}_N)$ and $\vec{p} = (\vec{p}_1, \ldots, \vec{p}_N)$ which are the coordinates, velocities and momenta off all particles.

Let us suppose that all the forces acting on the N–particle system are conservative; this means that it must exist a scalar function $U = U(\vec{r})$ called Potential Energy Function (PEF), such that

$$\vec{F}(\vec{r}) = -\partial_{r_i} U(\vec{r})\ \hat{e}_i = -\vec{\nabla}_{\vec{r}} U(\vec{r}) \tag{1.1.4}$$

where $\vec{F} = (\vec{F}_1, \ldots, \vec{F}_N)$. Thus we have only to know the PEF of the system at any time and the initial conditions to solve Newton's laws.

The kinetic energy of the system is defined as

$$K(\dot{\vec{r}}) = \sum_{i=0}^{N} \frac{1}{2} m_i \dot{\vec{r}}_i \cdot \dot{\vec{r}}_i \tag{1.1.5}$$

If the system is conservative we can define a scalar function, called *Lagrangian* of the system

$$\mathcal{L}(\vec{r}, \dot{\vec{r}}) = K(\dot{\vec{r}}) - U(\vec{r}) \tag{1.1.6}$$

such that

$$\frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{r}_i}\right) - \frac{\partial \mathcal{L}}{\partial r_i} = 0 \tag{1.1.7}$$

These set of 3N equations are called *Euler–Lagrange equations of motion*. It is easy to show that substituting the definition of $\mathcal{L}$ we obtain the Newton's second law. The Euler–Lagrange equations are a sort of generator of the equations of motion.

Using the definition of $\mathcal{L}$ (1.1.6) we have

$$p_i = \frac{\partial \mathcal{L}}{\partial \dot{r}_i} = m_i \dot{r}_i \tag{1.1.8}$$

thus we can express particle velocities as a function of particle momenta. Equations (1.1.5) and (1.1.8) let us to express the kinetic energy in the form

$$K(\vec{p}) = \sum_{i=0}^{N} \frac{\vec{p}_i \cdot \vec{p}_i}{2m_i} \tag{1.1.9}$$

To describe the system we can define another scalar function, called *Hamiltonian* of the system

$$\mathcal{H}(\vec{r}, \vec{p}) = \sum_{i=0}^{N} \vec{p}_i \cdot \dot{\vec{r}}_i(\vec{p}_i) - \mathcal{L}(\vec{r}, \dot{\vec{r}}(\vec{p}))$$

Substituting (1.1.6) and using (1.1.9) the Hamiltonian of the system is nothing that

$$\mathcal{H}(\vec{r}, \vec{p}) = K(\vec{p}) + U(\vec{r}) \tag{1.1.10}$$

or *the total energy of the system*. To obtain the equations of motion we have to solve *Hamilton's equations*

$$\begin{aligned} \dot{r}_i &= \frac{\partial \mathcal{H}}{\partial p_i} \\ \dot{p}_i &= -\frac{\partial \mathcal{H}}{\partial r_i} \end{aligned} \tag{1.1.11}$$

Describing the system with the Hamiltonian formalism, in some cases, is more useful than Lagrangian one, first of all because the Hamiltonian of a system is directly related to a well know physical quantity, the total energy.

1.2 REVIEW OF STATISTICAL MECHANICS

Using the picture of the classical mechanics described above we have a good and sophisticated machinery that allows us, knowing some information about the system in exam, i.e. initial positions and velocities of all particles and their interactions, to completely solve the equations of motion in order to get the dynamics of the system at every time. So classical mechanics encodes all the information about the *microscopic* view of a system and, in principle, we can extract all the information we want about the *macroscopic* proprieties of such system. The main task of this process is to obtain the thermodynamics proprieties of a system (temperature, pressure and so on) from the complete sets of positions and velocities of all particles and thus it is necessary to have a link between microscopic and macroscopic world. In principle this can be done, but if we consider a real system we should solve a set of 6N equations where N is of the order of the Avogadro number ($\mathcal{N}_A = 6.022 \cdot 10^{23}$ mol^{-1}); we can not think of solving such a number of

equations analytically even if we consider to solve it numerically: that is almost impossible. Thus the problem is to extract the macroscopic information from classical mechanics and to establish a well computable link between microscopic and macroscopic to obtain "easily" the thermodynamics information required.

The solution of that problem comes from *statistical mechanics* developed, principally, by Boltzmann and Gibbs. Statistical mechanics involves all the rules and methods through which the microscopic and macroscopic worlds are related to each other; this provides also a rigorous derivation of thermodynamics from the microscopic proprieties: without that thermodynamics would be only a phenomenological theory. The first step to the solution of the problem is to recognize that *a macroscopic observable of a system does not strongly depend on the complete dynamics of each particle in the system, but rather on an average that cancels out all the details of the microscopic features*. This it is intuitively true. We can think to set up an experiment with a system in a specific microscopic state that corresponds a macroscopic state. Certainly we can do the contrary and for sure we will not find the same microscopic state! Then we can iterate the experiment and we will find that for a specific macroscopic state of a system there exists a number of microscopic states that yield to the same properties.

The most important concept that makes this idea practicable is the *statistical ensemble*. Based on the previous considerations, a general definition of an ensemble is *a collection of systems subject to a set of common interactions and sharing the same macroscopic properties*. This concept lays the foundations of thermodynamics and suggests a procedure for computing many macroscopic observable. In more details a N–particle system in a specific microscopic state is described by its microstate $\vec{x}$, then *an ensemble is a set of points in the phase space that are subject to the constraint to be part of the ensemble itself*. Each system evolves in time with the equations of motion, so the time evolution of an ensemble is described by the flow of a set of points in the phase space according to the classical mechanics.

Once an ensemble is defined we are able to compute, at every time, the macroscopic observables simply averaging over all systems in the ensemble. To do this we have to know, at every time, which microstates of the phase space are part of the ensemble. For this purpose we define the *ensemble distribution function* $\tilde{\rho} = \tilde{\rho}(\vec{x}, t)$. Let $dx = dr_1 \cdots dr_{3N} dp_1 \cdots dp_{3N}$ the infinitesimal phase space volume, then

$$\frac{1}{\mathcal{N}} \tilde{\rho}(\vec{x}, t)\ dx = \rho(\vec{x}, t)\ dx????$$

where $\mathcal{N}$ is the total number of microstates in the ensemble. $\rho(\vec{x}, t)dx$ can be interpreted as the probability to find in the ensemble a system with microstate $\vec{x}$ at a time t and $\rho(\vec{x}, t)$ is the more convenient normalized distribution function. For definition of probability density must be

$$\int_\Omega \rho(\vec{x}, t)\ dx = 1, \qquad \rho(\vec{x}, t) \geqslant 0$$

Giving the ensemble distribution function, the ensemble average of an observable $A = A(\vec{x})$, at every time, is defined as

$$\langle A \rangle (t) = \int_{\Omega} A(\vec{x}) \rho(\vec{x}, t) \, dx$$

For an ensemble at thermodynamic equilibrium the macroscopic state is fixed and so, if A is an equilibrium observable, it must be time–independent. Thus, it can be shown [22] that exist a scalar function of the Hamiltonian $f(\mathcal{H}(\vec{x}))$ such that the time–independent ensemble average of the equilibrium observable A can be expressed as

$$\langle A \rangle = \frac{1}{\mathcal{Z}} \int_{\Omega} A(\vec{x}, t) f(\mathcal{H}(\vec{x})) \, dx$$

where $\mathcal{Z}$, known as *partition function*, is specific for the ensemble in exam and it is defined as follow

$$\mathcal{Z} = \int_{\Omega} f(\mathcal{H}(\vec{x})) \, dx$$

In order to compute the partition function we need to specify the thermodynamic observables, called *control variables*, that characterize the ensemble itself. By definition of an ensemble at thermodynamic equilibrium those control variables must be constant in time. The main ensembles used in statistical mechanics and the related control variables are summarized as follow

- *microcanonical ensemble*: constant–NVE
- *canonical ensemble*: constant–NVT
- *isothermal–isobaric ensemble*: constant–NpT
- *grand–canonical ensemble*: constant–μpT

The averages computed in different ensembles are equivalent in the so called *thermodynamic limit*: this is the *equivalence of ensembles*. Thus, it must be possible to change from one ensemble to another leaving averages unchanged.

We have defined ensemble averages and how to compute them but we need also a link between statistical averages and the experimental values. When we measure a macroscopic observable A we prepare an experiment with *only one* system in a specific macroscopic state and we study its evolution in time. A is a function of time and phase space vector and it fluctuates over time due to particle interactions. The measurement itself requires long time intervals compared to microscopic time scales, thus when we measure an observable we take an *average over time*. If, in principle, the time for average is infinity then we have the "real" mean value of the observable

$$\overline{A} = \lim_{\tau \to +\infty} \frac{1}{\tau} \int_{t_0}^{\tau} A(\vec{x}_t) \, dt$$

In order for a comparison to be made, an identity between ensemble and time averages must be established. This link is provided by the *ergodic theorem* and the *ergodic hypothesis*. A system is said to be ergodic if, over a long period of time, all

the microstates in the phase space with the same energy are accessible with the same probability. Then the ergodic theorem says that, if the system is ergodic, the time and ensemble averages are equal *almost everywhere* in the phase space. So we can write the follow identity

$$\overline{A} = \lim_{\tau \to +\infty} \frac{1}{\tau} \int_{t_0}^{\tau} A(\vec{x}(t))\, dt = \int_{\Omega} A(\vec{x}) \rho(\vec{x}, t)\, dx = \langle A \rangle (t) \tag{1.2.1}$$

For biomolecular applications the most important ensembles are the microcanonical and the isothermal–isobaric. In the following we will describe them briefly with particular attention to the isothermal–isobaric ensemble, the most relevant for this thesis work.

1.2.1 Microcanonical ensemble

The microcanonical ensemble is composed of the systems whose number of particles (N), volume (V) and energy (E) are constant. Due to the constant energy it describes a Hamiltonian system for which

$$\mathcal{H}(\vec{x}) = E$$

this let us to define the partition function as follow

$$\mathcal{Z}_{NVE} = \frac{1}{N! h^{3N}} \int_{\Omega} \delta(\mathcal{H}(\vec{x}) - E)\, dx \tag{1.2.2}$$

where the normalization factor N! takes into account the particles indistinguishability and h^{3N} is for compatibility of statistical mechanics with quantum mechanics: it comes from Heisenberg's uncertainty principle and it is the smallest phase space volume element. The right thermodynamic potential, i.e. the constant of motion, to obtain all the macroscopic observables is the entropy S given by

$$S = k_B \ln \mathcal{Z}_{NVE}$$

where $k_B = 1.3806505(24) \cdot 10^{-23}$ J/K is the Boltzmann constant. The other thermodynamic quantities can be obtained by

$$\frac{1}{T} = \left(\frac{\partial S}{\partial E}\right)_{NV} \qquad \frac{p}{T} = \left(\frac{\partial S}{\partial V}\right)_{NE} \qquad \frac{\mu}{T} = \left(\frac{\partial S}{\partial N}\right)_{VE}$$

The link between microscopic functions of phase space points and macroscopic observables, like kinetic energy or pressure, is provided by the *classical virial theorem* which states that

$$\left\langle x_i \frac{\partial \mathcal{H}}{\partial x_k} \right\rangle = k_B T \delta_{ik} \tag{1.2.3}$$

where x_i is some phase space variable.

KINETIC ENERGY Since a system in a microcanonical ensemble is defined to be Hamiltonian, from equations (1.1.10) and (1.1.9) according to the virial theorem with the choice $x_i = p_i$ we obtain

$$\left\langle \frac{p_i^2}{m_i} \right\rangle = k_b T$$

then summing both side over all particles and over the three coordinates we obtain the average kinetic energy at equilibrium

$$\langle K \rangle = \left\langle \sum_{i=1}^{N} \sum_{\alpha=1}^{3} \frac{p_{i+\alpha}^2}{2m_i} \right\rangle = \left\langle \sum_{i=1}^{N} \frac{\vec{p}_i \cdot \vec{p}_i}{2m_i} \right\rangle = \frac{3}{2} N k_B T \tag{1.2.4}$$

this is like the well know equipartition theorem in which 3N is the number of Degrees of Freedom (DOF).

PRESSURE Choosing $x_i = r_i$ in equation (1.2.3) and summing both side over all particles and substituting equations (1.1.11) and (1.2.4), we obtain

$$W = \left\langle \sum_{i=1}^{3N} r_i \dot{p}_i \right\rangle = -3 N k_B T = -2 \langle K \rangle$$

the quantity W is often called *virial*. For a N–particle non–interacting system it is well known that $pV = N k_B T$ thus the virial is $W = -3pV$. For a real interacting system we need to include in the virial the contribution due to the inter–particles interactions $U(\vec{r}_1, \dots, \vec{r}_N)$, thus the virial becomes[1]

$$W = -3pV + \left\langle \sum_{i=1}^{N} \sum_{j=i+1}^{N} r_{ij} f_{ij} \right\rangle = -2 \langle K \rangle$$

where r_{ij} and f_{ij} are the distance and force between particles i and j. Thus the pressure of the system is given by

$$p = \frac{1}{3V} \left(2 \langle K \rangle + \left\langle \sum_{i=1}^{N} \sum_{j=i+1}^{N} r_{ij} f_{ij} \right\rangle \right)$$

The instantaneous pressure in terms of the phase space points $\vec{x}(t)$ is obtained substituting equation (1.2.4) and getting the quantity in the average bracket, so

$$p(\vec{x}_t) = \frac{1}{3V} \sum_{i=1}^{N} \left(\frac{\vec{p}_i \cdot \vec{p}_i}{m_i} + \sum_{j=i+1}^{N} r_{ij} f_{ij} \right) \tag{1.2.5}$$

1 We assume that the inter–particles interactions are pairwise addictive which they depends only on distances between particles i and j.

1.2.2 Isothermal–isobaric ensemble

The isothermal–isobaric ensemble contains those systems with constant particles number (N), pressure (p) and temperature (T). This is useful in many chemical, biological and physical systems since their properties are reported in conditions of standard temperature and pressure. To maintain the system at constant temperature and pressure it is necessary to couple it with an external *temperature bath* and a *pressure bath*. The first one can be considered simply a very big system at constant temperature with a high thermal capacity. The second can be idealized like a piston connected to the system: it change the volume in order to adjust the pressure. The instantaneous work done by the system against the external piston is defined by pV, where V is the instantaneous system volume. Then we have to correct the Hamiltonian of the system: $\mathcal{H}(\vec{x}) \to \mathcal{H}(\vec{x}) + pV$. The partition function is then defined considering the Boltzmann ensemble distribution function

$$\mathcal{Z}_{NPT} = \frac{1}{N!h^{3N}} \int_0^{+\infty} dV \int_\Omega e^{-\beta(\mathcal{H}(\vec{x})+pV)}\, dx \tag{1.2.6}$$

where $\beta^{-1} \equiv k_B T$ and the normalization factor is the same as in equation (1.2.2). The right thermodynamic potential, i.e. the constant of motion, to obtain the other thermodynamic quantities is the Gibbs free energy $G = H - TS$ (where H is the enthalpy) defined by

$$G = -k_B T \ln \mathcal{Z}_{NpT}$$

it describes the maximum reversible work that may be performed by the system. The other thermodynamic quantities can be obtained by

$$\mu = \left(\frac{\partial G}{\partial N}\right)_{pT} \qquad \langle V \rangle = \left(\frac{\partial G}{\partial p}\right)_{NT} \qquad S = \left(\frac{\partial G}{\partial T}\right)_{Np}$$

For anisotropic systems, volume can undergo different changes in the three Cartesian directories even if external pressure is applied isotropically. In these cases it is necessary to take this into account in the partition function. The way is to define a matrix formed by the three basis vectors of the system box $\vec{a}$, $\vec{b}$ and $\vec{c}$

$$\mathbf{H} = \begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{pmatrix} \tag{1.2.7}$$

so that its volume is $V = \vec{a} \cdot \vec{b} \times \vec{c} = |\det \mathbf{H}|$. The partition function become

$$Z_{NpT} = \frac{1}{N!h^{3N}} \int d\mathbf{H} \int_\Omega \frac{1}{(\det H)^2} e^{-\beta(\mathcal{H}(\vec{x})+p|\det \mathbf{H}|)}\, dx$$

where $\int d\mathbf{H}$ is an integral over all nine components of $\mathbf{H}$. Hence, the instantaneous pressure can no longer be described by a single quantity. Instead a 3×3 pressure

matrix $\mathbf{P}$ is needed. What one find, if the system is isotropically coupled, is that on average, this pressure matrix reduces to a diagonal pressure matrix such that

$$\mathrm{Tr}\,\langle \mathbf{P} \rangle = 3p$$

1.3 MOLECULAR DYNAMICS SIMULATION

Molecular Dynamics (MD) is a set of techniques that allow us to prepare a "computer experiments" in which solving numerically the classical equations of motion of a virtual system we will be able to know its time evolution. Such virtual experiment approach has the advantage that many experiments can be set up with different initial conditions and/or with different control parameters, such as temperature or pressure. Obviously that experiment is carried out using a model that approximates the real system. The main parts of that model are the information required to obtain an approximation of the interactions among all system particles, i.e. to compute the PEF from which the forces are derived by equation (1.1.4). Solving the equations of motion with a numerical integrator, an MD simulation generates a set of phase space vectors, a *trajectory*, at discrete times that are multiples of the fundamental time discretization parameter, called *MD time step*; δt. Starting from an initial phase space vector $\vec{x}(0)$, at each step, the forces are computed from the PEF. Then the equations of motion are integrated and a new phase space vector $\vec{x}(\delta t)$ is generated, thus a new set of forces is computed and so on. In order to compute time averages we need to discretize equation (1.2.1) so the time integration is substituted with a summation over the collected data at certain time step $\Delta\tau = i\delta t$, $i = 1, 2, 3, \ldots$. If $i > 1$ only a subset of the collected data is used to compute time averages. The formula becomes

$$\langle A \rangle = \frac{1}{M} \sum_{n=1}^{M} A(\vec{x}(n\Delta\tau)) \tag{1.3.1}$$

where $M\Delta\tau$ is the total averaged time and of course it must be $M \leqslant D/i$ if D is the total number of MD steps. An MD simulation can be summarized in the scheme in figure (1).

1.3.1 Initial configuration

Before an MD simulation can be performed it is necessary to select an *initial configuration*. Its choice can be nontrivial and it depends on the complexity of the system. Then, careful attention must be paid in setting up the initial configuration.

Setting up an initial configuration means to prepare an N–particle system and assign all particle positions and velocities, i.e. all the $6N$ coordinates of the initial phase space vector $\vec{x}(0)$. A common choice to assign the initial velocities is to

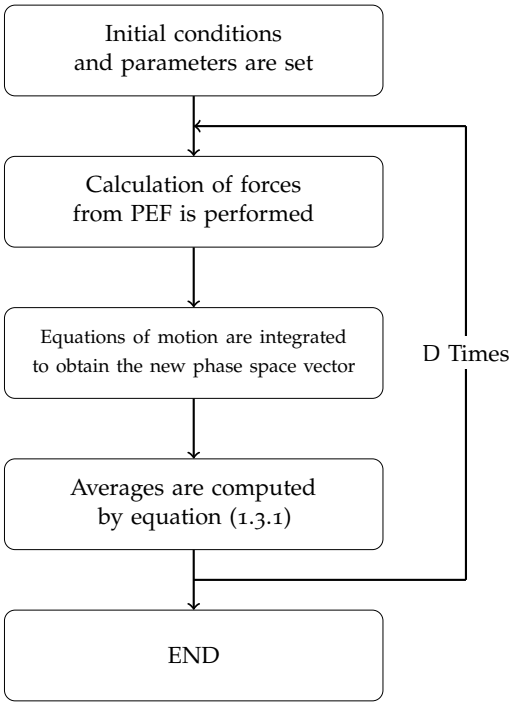


Figure 1: Schematic representation of an MD simulation.

extract they randomly from the Maxwell–Boltzmann distribution function at a specific system's temperature

$$f(v_i) = \sqrt{\frac{m_i}{2\pi k_B T}} e^{-\left(\frac{m_i v_i^2}{2k_B T}\right)}$$

Moreover, the random assignment algorithm has to rescale all the velocities in such a way that the total system's momentum $\vec{P} = \sum_{k=1}^{N} m_k \vec{v}_k$ is zero, this is equivalent to a Center of Mass (COM) motion removal. This is done because, in general, the total force acting on the system $\vec{F} = \sum_{k=1}^{N} \vec{F}_i$ is zero, then the COM motion is constant and to avoid a constant drift of the system in space this can be removed. Of course this is a constraint on the system and it must be taken into account because it reduces the system DOF by three.

1.3.2 Periodic boundary conditions

In an MD simulation the sample system is inserted into a *simulation box* whose shape can be differently chosen to better reproduce the symmetry of the simulated system. That box gives us the trivial possibility to introduce a well defined reference system of coordinates. Obviously we must not forget to correctly treat the *boundary conditions*. In order to avoid surface effects and to consider only an infinite bulk system, Periodic Boundary Conditions (PBC) are imposed to the simulation box. This gives us also the possibility to simulate system's bulk proprieties without considering a too large number of particles. To give a better idea, in fig-

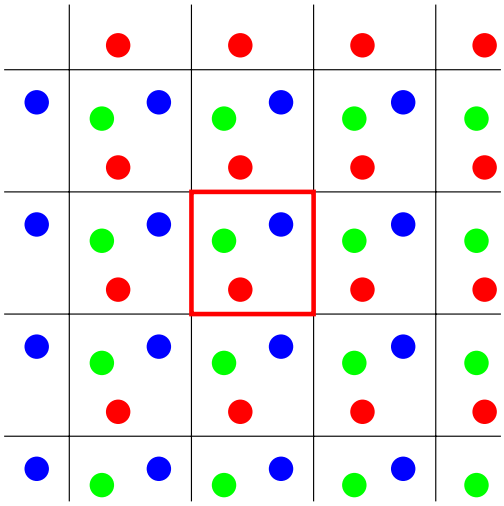


Figure 2: Schematic view of a two–dimensional box with PBC imposed. The central, red contoured, box is the simulation box and it is replicated along each side.

ure (2) an example of a two–dimensional box with PBC is shown. The central red contoured box is the simulation box. The idea is to replicate that box in space along each side so that there are no surface particles nor walls in the central box. When a particle moves in the central box, all its images virtually move the same way in the copies of the box so that if a particle leaves the virtual boundary of the central box, then, its nearest image enters the box and the number density of particles in the simulation box is conserved. This virtual movement of image particles is achieved adjusting the positions of the simulation box particles which have left the main box. For example, if one use a cubic box and a particle crosses its boundary in one direction, say the x direction, then its coordinate is corrected by subtracting (if it leaves the box in the positive direction) or adding the box side length parallel to x direction. Only box geometries compatible with translational symmetry can be used. For example, nether a spherical nor a icosahedral box could do the job. However when it is possible one have to use the most appropriate shape to better describe the symmetry of the system, otherwise a closest approximation, compatible with PBC, must be used.

Even if PBC are used in a wide range of applications, it must be taken into account that imposing periodicity to a system may affect its properties. A clear limitation of the periodic cell is that it is not possible to achieve fluctuations that have a wavelength greater then the cell length. This cause, obviously, the impossibility to sample those vibrating modes. Another problem arises with the range of the inter–particles interactions: one have to choose carefully the size of the simulation box, or the number of particles if an NPT ensemble is used, to ensure that the smallest simulation box length is greater then the interaction range. This can be made easily for example with the Van der Waals interaction. On the contrary it is a difficult and time consuming task to do the same with the electrostatic interactions that are treated with a more sophisticated methods, as better explained in section 1.4.4.

1.3.3 Numerical integrators

As we have seen above we need to solve numerically the equations of motion. Since the PEF is a continuous function of the phase space vector at a time t, the simplest way is to use the so called *finite difference* method. The basic idea is to expand the Newton's law in a Taylor series as follow

$$\vec{r}_i(t+\delta t) = \vec{r}_i(t) + \vec{v}_i(t)\ \delta t + \frac{1}{2m_i}\vec{f}_i(t)\ (\delta t)^2\ +\ o((\delta t)^3) \tag{1.3.2}$$

where we used the identities $\vec{v}_i(t) = \dot{\vec{r}}_i(t)$ and $m_i\ddot{\vec{r}}_i(t) = \vec{f}_i(t)$.

From this point, different algorithms have been developed. In the following we will describe in detail the most important, the *Verlet algorithm*, and its implementation the *leap–frog algorithm*, which is the default used in our MD tools for this thesis work.

Verlet algorithm

The Verlet algorithm requires the positions and the forces at a time t and the positions at a time $t-\delta t$ to calculate the positions at a time $t+\delta t$. Starting from equation (1.3.2) we can write

$$\vec{r}_i(t+\delta t) \simeq \vec{r}_i(t) + \vec{v}_i(t)\delta t + \frac{1}{2m_i}\vec{f}_i(t)\ (\delta t)^2 \tag{1.3.3}$$

$$\vec{r}_i(t-\delta t) \simeq \vec{r}_i(t) - \vec{v}_i(t)\delta t + \frac{1}{2m_i}\vec{f}_i(t)\ (\delta t)^2 \tag{1.3.4}$$

from their sum we obtain the new positions at a time $t+\delta t$

$$\vec{r}_i(t+\delta t) \simeq 2\vec{r}_i(t) - \vec{r}_i(t-\delta t) + \frac{1}{m_i}\vec{f}_i(t)\ (\delta t)^2 \tag{1.3.5}$$

The velocities do not appear in the equation above and can be obtained taking the difference of equation (1.3.3) and (1.3.4)

$$\vec{v}_i(t) \simeq \frac{\vec{r}_i(t+\delta t) - \vec{r}_i(t-\delta t)}{2\delta t}$$

Since positions in equation (1.3.5) are computed as differences this is a fourth order algorithm and the precision is up to $o(\delta t)^4$, but they also contain a small term of order $o(\delta t)^2$, $(\vec{f}_i(t)/(2m_i))$ which is summed to a difference of larger terms $(2\vec{r}_i(t) - \vec{r}_i(t-\delta t))$. This may cause a loss of precision due to computer numerical representation.

The main disadvantage is that velocities at a time t are an output of the calculation and not a part of the algorithm itself. Moreover it is not self–starting because the algorithm required the positions at a time $t-\delta t$. So at $t = 0$ we need a trick to obtain the previous inexistent positions. The trick is to use equation (1.3.4) truncated at the first order: $\vec{r}_i(-\delta t) \simeq \vec{r}(0) - \vec{v}_i(0)\delta t$.

Leap–Frog algorithm

The leap–frog algorithm is a variant of the Verlet one and it is commonly implemented in many MD tools, as it is in our case. It computes the positions at a time t and the velocities at a time $t + 1/2\delta t$ from the forces at a time t and the velocities at a time $t - 1/2\delta t$. The main advantage with respect to the Verlet algorithm, is that it is self–starting because it does not require the positions at a time $t - \delta t$.

First it calculates the velocities at a time $t + 1/2\delta t$ as follow

$$\vec{v}_i(t + 1/2\delta t) \simeq \vec{v}_i(t - 1/2\delta t) + \vec{a}_i(t)\delta t$$

then the positions at a time $t + \delta t$ are computed

$$\vec{r}_i(t + \delta t) \simeq \vec{r}_i(t) + \vec{v}_i(t + 1/2\delta t)\delta t$$

The velocities at a time t can be calculated by

$$\vec{v}_i(t) \simeq \frac{\vec{v}_i(t + 1/2\delta t) + \vec{v}_i(t - 1/2\delta t)}{2} \tag{1.3.6}$$

Another advantage is that the velocities are part of the algorithm itself and moreover it does not require the calculation of the difference between two large numbers, with a gain of precision. The obvious disadvantage is that the positions and velocities are not synchronized so the equation (1.3.6) is necessary to calculate the velocities at a time t. The need to have velocities at the same time of positions, as for the Verlet algorithm, derives from the calculation of the kinetic energy contribution to the total energy: it must be computed with positions and velocities at the same time.

1.3.4 Neighbor list

In order to solve the classical equations of motion, it is necessary to know the forces, and so the PEF, acting on the system's particles. As we shall see in detail in the next section, this is one of the most time consuming part of an MD simulation. To know the forces acting on one particle, in principle, it is necessary to calculate the contribution from all particles in the simulation box and all other periodic images. The most popular way to speed up the simulation is to use a truncation of the interaction potentials within a cutoff. The general idea is to compute the energy contribution of particle i considering the interaction only with particles j that are closer to a certain cutoff distance r_c, thus such that $|\vec{r}_i - \vec{r}_j| \leqslant r_c$. This is summarized in the following expressions

$$U_i(\vec{r}) = \sum_{j=0}^{N} U^*_{ij}(\vec{r}), \qquad U^*_{ij}(\vec{r}) = \begin{cases} U_{ij}(\vec{r}) & |\vec{r}_i - \vec{r}_j| \leqslant r_c \\ 0 & \text{otherwise} \end{cases}$$

where $U_{ij}(\vec{r})$ is the interaction potential between particles i and j.

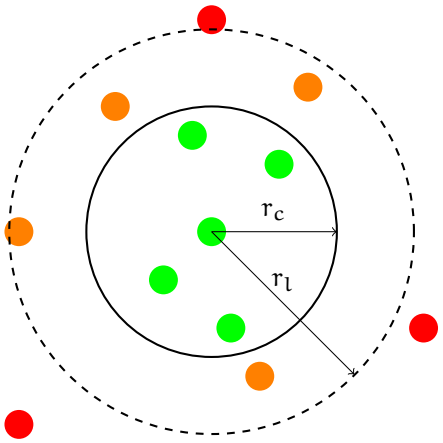


Figure 3: Schematic representation of the buffered pair–list construction respect to the central particle. Green particles are in the pair–list below the cut-off radius r_c, therefore included in the calculation of the interactions. Orange particles are in the pair–list for which at every step is checked if their distances became smaller then r_c. Red particles are not in the pair–list and they are completely neglected until the next list update.

By itself, the use of a truncation of the potential may not dramatically reduce the time spent in computing the inter–particle interactions. This is because, in order to decide for what particles we have to compute the interactions, we have to compute the distances between every pair of particles in the system. A marked increase of performance is achieved by the use of a *neighbor pair–list*. The simplest way is to consider, for each particle, a list of its neighbor particles that lie within a sphere of radius r_c surrounding the selected particle. Then, for each particle, the interactions are computed between the selected particle and those that are in its pair–list. There is a gain in performance if the pair–list is updated at least every $M > 1$ MD steps. This is possible taking into account that, typically, in MD simulations of liquids system at ambient temperature and pressure, a particle neighbors do not change significantly for $M < 20$ time steps.

Anyway particles move during the non–updating time so that some of them may cross the pair–list causing an over– or under–estimation of the inter–particle energy contribution. To partially solve this problem, as suggested by Verlet [23], one can consider a *buffered* pair–list or a *Verlet cut–off scheme* in which the pair–list is constructed considering those particles that are close to the selected one by a distance of $r_l > r_c$, called *list radius*. That pair–list is updated every some time steps, but every time steps the pairwise contribution is computed only between those particles in the pair–list for which the distance is less then r_c. Thus, at the cost of slightly decrease the performance compared to the un–buffered pair–list, almost none of the interacting particles within the cutoff is neglected. Nevertheless, since a too big list radius and a too small list update frequency leads to a loss of performance, particle–pairs could still move enough during the non–updating time and have a chance to cross the boundary of the buffered pair–list r_l. This chance leads to a small energy drift proportional to the system temperature. In simulations with a constant temperature coupling, i.e. in a canonical or isothermal–isobaric ensemble, the extra radius of the buffered pair–list can be dynamically determined, during the simulation, by fixing an upper limit to the energy drift. That tolerance is often called *Verlet buffer tolerance*. In figure (3) a schematic representation of the buffered pair–list construction is shown.

A better performance can be achieved adding a dynamic update algorithm of the pair–list refresh rate during the simulation. A nice way is to consider the maximum distances traveled by a particles in the pair–list: if this distance is greater than $r_b = r_l - r_c$ then the pair–list is certainly updated. From this one can fix the refresh rate to a higher value increasing the performance.

1.3.5 Thermostat algorithms

A thermostat is an external tool that allows to maintain a system at a constant temperature. Several algorithms are available. Some are based acting on particle velocities (Anderson, Berendsen, Bussi) other introduce some more DOF in the system that take into account a real temperature bath coupling (Nosé–Hoover). We will describe in more detail the one used in this thesis work.

As suggested by Bussi *et al.* [3], a common practice to implement a thermostat acting on particle velocities is related to a *velocity rescale* algorithm in which the velocities of all particles are scaled by some factor. The simplest way is to consider the total kinetic energy K of the system as in equation (1.1.5) and the average kinetic energy $\langle K \rangle$ obtained from equation (1.2.4) with the substitution $3N \to N_f$

$$\langle K \rangle = \frac{1}{2} N_f k_B T$$

where N_f is the *total* DOF of the system and T is the target temperature. Thus the scaling factor is defined by

$$\alpha_T \equiv \sqrt{\frac{\langle K \rangle}{K}}$$

The scaling operation is usually performed at a fixed rate during the simulation, or when the kinetic energy exceeds the limits of an interval centered around the target value. However the sampled ensemble is not explicitly known but, since in the thermodynamic limit the average properties do not depend on the ensemble chosen, even this very simple algorithm, often called *weak coupling thermostat*, can be used to produce useful results. Despite this, for small systems or when the observables of interest are dependent on the fluctuations rather than on the averages or when other methods assume a canonical sampling, this method cannot be safely used.

In order to obtain the correct canonical sampling Bussi *et al.* [3] modify the way to calculate α_T so as to enforce a canonical distribution for the kinetic energy. The new scaling factor is obtained from

$$\alpha_T \equiv \sqrt{\frac{K_T}{K}}$$

where K_T is extracted with a stochastic procedure from the equilibrium canonical distribution of the kinetic energy, given by

$$P(K_T) dK_T \propto K_T^{N_f/2-1} \, e^{-\beta K_T} \, dK_T$$

where $P(K_T)dK_T$ is probability that the system has a kinetic energy between K_T and $K_T + dK_T$.

Since velocities are scaled only after some MD steps this can cause a discontinuity in the particles velocities just before and after the scaling step. To avoid this problem, the authors suggest that the choice of K_T can be based on the previous value of K so as to obtain a smoother evolution. The procedure proposed by Bussi *et al.* consists in the following steps

- Evolve the system for a single time step solving the equations of motion, so as to be in an NVE ensemble;
- Calculate the total kinetic energy and evolve it for a single time step using an auxiliary continuous stochastic dynamics;
- Rescale the velocities by α_T so as to enforce this new value of the kinetic energy.

The authors have shown that an auxiliary dynamics for the kinetic energy of the form

$$dK = \frac{\langle K \rangle - K}{\tau_T}\, dt + 2\sqrt{\frac{K \langle K \rangle}{N_f \tau_T}}\, dW$$

can do the job. Where dW is a Wiener stochastic noise and τ_T is an arbitrary parameter which is related to the response time of the thermostat. In fact it reduces to the simple week coupling rescale if $\tau_T \to 0$ and to the Hamiltonian dynamics (as an NVE ensemble) if $\tau_T \to +\infty$. Until the system is at non–equilibrium the first deterministic part is dominant an drives the system to equilibrium with a characteristic time τ_T. Then the stochastic contribution samples the canonical distribution.

1.3.6 Barostat algorithms

As the thermostat maintains the system at a constant temperature T, a barostat is needed to maintain the system at a constant pressure p. To do this the system is coupled to a sort of piston so that changing the volume of the simulation box will adjust the pressure.

Berendsen algorithm

A common way, as proposed by Berendsen, is to scale both volume and particle coordinates. The rate of change of the pressure is given by

$$\frac{dp}{dt} = \frac{p_0 - p}{\tau_p}$$

where p_0 is the target pressure, p is the instantaneous system pressure given by equation (1.2.5) and τ_p is a coupling constant related to the response time of the

barostat. Thus if we consider a time step δt then the volume scaling factor λ is given by

$$\lambda = 1 - k_T(p_0 - p)\frac{\delta t}{\tau_p}$$

where k_t is the isothermal compressibility defined as

$$k_T = -\frac{1}{V}\left(\frac{\partial V}{\partial p}\right)_T$$

while the factor $\delta t/\tau_p$ gives a scaling factor for the isothermal compressibility that allow us to take into account the finite response time of the barostat. If $\tau_p \to 0$ the system has an infinity isothermal compressibility so it is necessary a really small change in volume to achieve the correct pressure; on the contrary if $\tau_p \to +\infty$ the system reduces to the Hamiltonian dynamics[2]. The particle coordinates is scaled by the factor $\lambda^{1/3}$.

In the case of anisotropic system the pressure matrix $\mathbf{P}$ and the volume matrix $\mathbf{H}$ have to be considered and the Berendsen algorithm can be generalized such that even λ becomes a 3×3 matrix. However the main problem, as for the weak coupling thermostat, is that the sampled ensemble is not known. Thus a new approach, in order to correct sample the isobaric ensemble, has been derived by Parrinello and Rahman.

Parrinello–Rahman algorithm

The Parrinello–Rahman barostat [18][19] is based on genuinely treat the coupling of the system to an external piston of "mass" M_h through a Lagrangian method in which the volume becomes a Lagrangian variable of the system and its time evolution can be computed solving the Euler–Lagrange equations. Both the size and the shape of the simulation box are allowed to fluctuate. So it is perfectly compatible with anisotropic system. The shape and size of the simulation box are described by the volume matrix $\mathbf{H}$ in equation (1.2.7) such that $V = |\det \mathbf{H}|$. $\mathbf{H}$ is the Lagrangian coordinate of the external piston. If the external pressure p_0 is applied to the piston then its potential energy is given by

$$U_p = p_0 |\det \mathbf{H}|$$

instead, if $\mathbf{H}$ varies in time then a "kinetic energy" is associated to the piston as follow

$$K_p = \frac{1}{2} M_h \mathrm{Tr}({}^t\dot{\mathbf{H}}\dot{\mathbf{H}})$$

where ${}^t(\cdot)$ denote the transpose operation.

In order to write the Lagrangian of the system the particle coordinates must be expressed in terms of $\mathbf{H}$. This can be done defining the vector $\vec{s}_i$ so that $\vec{r}_i = \mathbf{H}\vec{s}_i$.

2 If the system is coupled to a thermostat, then the canonical ensemble is sampled.

The square displacement can be obtained as $\vec{r}_i \cdot \vec{r}_i = {}^t(\mathbf{H}\vec{s}_i)\mathbf{H}\vec{s}_i = {}^t\vec{s}_i\,{}^t\mathbf{H}\mathbf{H}\vec{s}_i$, often $\mathbf{G} = {}^t\mathbf{H}\mathbf{H}$. Parrinello *et al.* write the Lagrangian as follow

$$\mathcal{L} = \frac{1}{2}\sum_{i=1}^{N} m_i \dot{\vec{s}}_i\,{}^t\mathbf{G}\dot{\vec{s}}_i - \sum_{i=1}^{N}\sum_{j=i+1}^{N} U(r_{ij}) + \frac{1}{2}M_h \mathrm{Tr}({}^t\dot{\mathbf{H}}\dot{\mathbf{H}}) - p_0|\det\mathbf{H}|$$

thus the equations of motion for $\mathbf{H}$ and $\vec{s}_i$ can be computed solving the Euler–Lagrange equations. It can be shown that those equations of motion, derived in [18] and [19], correct sample an isobaric ensemble. A practical way to show this is to consider the Hamiltonian of the system. Using equation (1.1.10) the Hamiltonian is

$$\mathcal{H} = \frac{1}{2}\sum_{i=1}^{N} m_i \dot{\vec{s}}_i\,{}^t\mathbf{G}\dot{\vec{s}}_i + \sum_{i=1}^{N}\sum_{j=i+1}^{N} U(r_{ij}) + \frac{1}{2}M_h \mathrm{Tr}({}^t\dot{\mathbf{H}}\dot{\mathbf{H}}) + p_0|\det\mathbf{H}|$$

if the system is in equilibrium at temperature T, using the equipartition theorem, the kinetic term of the system particles contributes to the energy by $3Nk_BT/2$ while the kinetic term of the piston by $9k_BT/2$. Thus since $3N \gg 9$ the Hamiltonian can be approximated by

$$\mathcal{H} \simeq \langle K\rangle + U + p_0V = H$$

that is the enthalpy. Since the Hamiltonian is a constant of motion the correct isobaric ensemble is sampled. If the system is also coupled to a thermostat, the $-TS$ term must be added, so the constant of motion become the Gibbs free energy $G = H - TS$ and the isobaric–isotherm ensemble is correctly sampled.

In order to understand the meaning of the M_h parameter (the piston "mass") we report the dynamic equation of $\mathbf{H}$

$$M_h\ddot{\mathbf{H}} = V(p - p_0)({}^t\mathbf{H})^{-1}$$

The Parrinello–Rahman barostat is like a second order system. When there is an imbalance between the instantaneous internal pressure p (see equation (1.2.5)) and the target pressure p_0, the system recovers this imbalance in a characteristic time governed by the parameter M_h. Since at equilibrium the properties of a system are independent of the masses of its constituent parts, M_h can be arbitrarily chosen if one is interested only in static averages, otherwise a more appropriate choice must be made to obtain accurate dynamical properties. In this case, the authors suggest a choice of M_h such that the relaxation time is of the order of L/c where L is simulation box size and c is the sound velocity.

Interesting is the fact that this algorithm can be generalized to a anisotropic pressure coupling making use of the theory of elasticity. Differently from the Berendsen algorithm, Parrinello–Rahman method is much slower in changing the volume until the equilibrium value is reached and it is less stable: if the system is not well equilibrated can lead to a large volume fluctuation with can compromise the simulation success. On the other hand the Parrinello–Rahman samples correctly the isobaric ensemble. A common strategy is to use the Berendsen barostat

to reach equilibrium then switch to the Parrinello–Rahman algorithm to correctly sample the phase space associated to the isobaric ensemble.

1.4 EMPIRICAL FORCE–FIELD MODEL

As we have seen in the previous section MD provides a variety of tools for solving the time evolution of a N–particles system to obtain its dynamics. Due to the possibility to capture different length and time scales MD simulations can be used in a variety of systems, such as set of atoms, molecules or more complex system such as protein and macromolecules systems. In each of these systems, depending on the *interaction model* and its *parametrization*, we will be able to describe crucial molecular–level processes, such as hydrogen bond formation in organic molecules, which happen on the picoseconds time scale; or study slow processes such as the diffusion of massive colloidal particles, taking place on time scales of milliseconds if not seconds. When we study a soft or condensed matter system or, in general, a system composed by a large number of atoms (of the order of $N \gg 1000$), a crucial role is played by the *Born–Oppenheimer approximation*. It says that we can separate the motion of the electrons by the motion of the atomic nuclei. That is done for integrating out the high frequency electrons' motions in order to remove some DOF. Moreover the main interesting processes of soft and condensed matter, ranging from protein folding to glass transitions, from surface diffusion to ligand–receptor binding take place on longer time scales and involve larger number of atoms. Further if we want to know precisely the dynamics of the electrons in the system we have to introduce quantum mechanical methods that are, even for a small number of particles (of the order of $N \sim 100$), to much computationally time consuming, thus the Born–Oppenheimer approximation is indispensable. In the following, when we speak about atoms or chemical moieties we refer to it as for nuclei coordinates only without considering electrons at all.

Nevertheless atoms or molecules interactions, such as bond formation, are mediated by the electrons interactions. Thus, to describe the dynamics of such a system with a classical MD tools and the Born–Oppenheimer approximation, it is necessary to develop an *empirical model of the inter–atoms interactions* that mimic correctly the "real" interactions. Since forces are derived from the PEF we need a model composed by the set of the simplest pairwise addictive potentials that mimic the inter–particles interactions. The model, the set of simulation parameters, such as the time step, the set of functional forms of the inter–particles interactions potential and its parameterizations are collected into the so called empirical Force Field (FF). The meaning of *empirical* is that most of the functional forms of the inter–atoms interaction has no "first principle" justification and they are only an approximation to reality: There is not a correct expression of them and they are chosen as a compromise between accuracy and computational efficiency. Further it is necessary to stress out that a FF is a well defined single entity containing the simulation parameters, the functional models of the interactions and also its parameterizations (and the way to obtain it). All the parameters of a FF are in

harmony to each other thus changing some parameters without retesting whole FF is not allowed because, maybe, one can destroy the whole FF.

For biomolecular applications two main classes of FFs exist: The *atomistic* FFs in which basic particles are atoms, and the *coarse–grained* FFs in which the basic particles represent atom groups or small chemical moieties. In this case, even the way to do the corse–graining of the atoms in the molecules, called *mapping*, is part of the FF itself. Different Coarse–Grained (CG) FFs can use different mapping methods even with the same functional forms. In the following we will add some other information about FFs and describe the principal functional forms for modeling the inter–particles interactions and how to treat them in a MD simulation. While in the next section we will focus on the main CG FF used in this thesis work: The MARTINI CG FF developed by Marrink *et al.* [15].

PARAMETERIZATION In general the functional forms for potential interactions are common to all particles in the system, then the FF is completed by a set of empirical parameters that characterize the interaction between different types of particles, whether they are atoms or whole chemical groups. Interaction parameters are empirical in the sense that they are assigned to reproduce a small set of target properties on a small group of systems. These target properties can be derived from experimental measurements or from finer–level calculations or simulations. Nowadays, atomistic and CG biomolecular FFs come as "packages" of parameters and functional forms appropriate for the description of a large variety of chemical compounds in the liquid and solid phases.

TRANSFERABILITY As described above the parameterization of a FF involves a small set of test systems for which some set of target proprieties are reproduced. The main characteristic of a FF is the *transferability* that means the ability of the model to describe different situations that differ from those used at the parameterization stage. Of course one would expect to be able to make some predictions for a bigger variety of systems and for other proprieties not used in the parametrization stage. Common faults of organic FFs concern, for example, phase transitions of organic compounds and phase transitions temperatures.

1.4.1 Inter–particles interactions

For biomolecular applications the inter–particles interaction potentials are divided into two main classes: The *bonded interactions* involving particles within the same molecules and the *non–bonded interactions* engaging all particles in the system and

which usually represent the Van der Waals and the electrostatic interactions. The most common and general functional form for the PEF is the following one

$$\begin{aligned} U(\vec{r}_1, \cdots \vec{r}_N) = & \frac{1}{2} \sum_{\text{bonds}} \frac{1}{2} k_i^b (l_i - l_{i0})^2 + \frac{1}{2} \sum_{\text{angles}} k_i^a (\theta_i - \theta_{i0})^2 + \\ & + \frac{1}{2} \sum_{\text{torsions}} V_n (1 + \cos(n\omega - \gamma)) + \\ & + \sum_{i=1}^{N} \sum_{j>i} \left(4\epsilon_{ij} \left(\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right) + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right) \end{aligned} \tag{1.4.1}$$

The first two terms in equation (1.4.1) are harmonic potentials which model respectively the energy contribution due to deviation from reference bond length l_{i0} and bond angle θ_{i0}. Together with the bond and angle elastic constants, k_{bi} and k_{ia} respectively, they constitutes the set of parameters for bond and angle contributions. The angle contribution involves a set of three particles in the same molecule. The middle line of equation (1.4.1) concerns the energy contribution due to the bond torsional change where ω is the torsional angle. It involves four particles in the same molecule and mimic the energy barrier needed to rotate the bond angle along the bond axis. γ is a phase factor, V_n qualitatively describes the energy barrier for each n–th components and n is defined as the number of minima for each components. The last line in equation (1.4.1) contains the energy contribution due to the non–bonded interactions: the Van der Waals modeled by a Lennard–Jones $12-6$ potential, fully characterized by the constants σ_{ij} and ϵ_{ij} proper for each particles pair; and the electrostatic potential described by the particles charge q. The non–bonded interactions involve obviously all particles in the system, but for particles belonging to same molecule they are computed only if they are separated by at least three bonds, i.e. if their interactions are not described by bonded terms. The various contributions described above are schematically represented in figure (4).

1.4.2 Non–bonded interactions

The bonded interactions, as we can see in equation (1.4.1), are at *fixed range*, meaning that they depend, for example, on the equilibrium bond length that is fixed. The same does not hold for the non–bonded interactions because they depend on the inter–particles distance r_{ij} and they decay to zero as a power of r_{ij}^{-d}. Depending on the power order d compared to the dimensionality s of the system they are split into *short range* if $d > s$ and *long range* interactions if $1 \leqslant d < s$. For example, as we shall see later, the Lennard–Jones $12-6$ potential decays to zero as r^{-6} then it is a short range interaction, while the electrostatic is a long range interaction since it decays to zero as r.

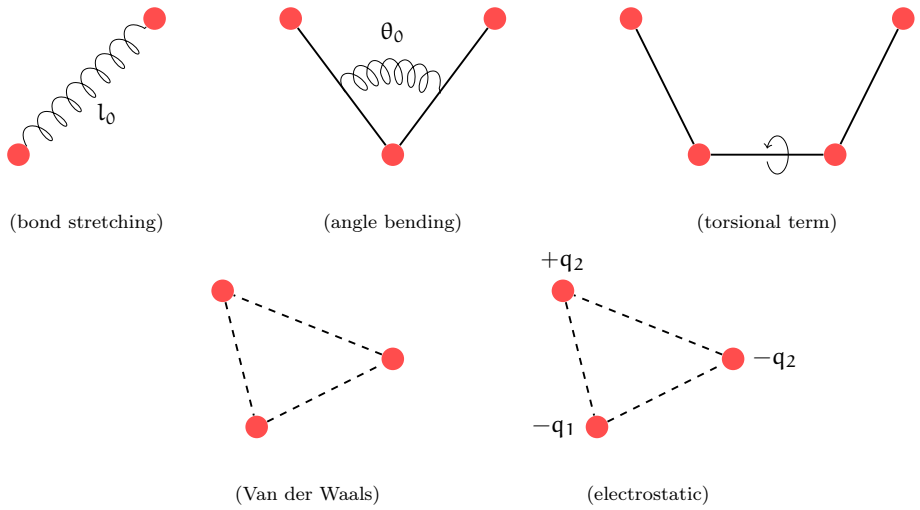


Figure 4: Schematic representation of the common inter–atoms interactions for biomolecular applications: bond stretching, angle bending, torsional term, Van der Waals and electrostatic interactions.

Cut–off, shift and switch methods

As we have mentioned in 1.3.4 the calculations of the non–bonded interactions energy contributions is one of the most time consuming part of an MD simulations. Since they are pairwise interactions their calculation scale as $\sim N^2$. Especially for the short range interactions, various methods were developed in order to speed–up the simulations. The *cut–off* method is the most used to treat the short range interactions and, in same cases, even the long range one. Obviously this is strictly connected to the way neighbor list are computed as we have seen a cut–off method alone is not enough to speed up the simulation. Taking one particle into account, the general idea is to evaluate the non–bonded interactions with all other particles that are closer to the first for a distance r_c, called *cut–off* radius, otherwise the interactions is set to 0. This means that the new potential is of the form

$$v^*(r) = \begin{cases} v(r) & r \leqslant r_c \\ 0 & r > r_c \end{cases}$$

This generates a discontinuity in the potential and in its first derivative, i.e. in the forces: This is bad for energy conservation. A trick for solving the discontinuity of the potential and to improve the energy conservation is to apply also a *shift* of the potential value at r_c, since it is a constant it does not affect the forces. We have

$$v^*(r) = \begin{cases} v(r) - v(r_c) & r \leqslant r_c \\ 0 & r > r_c \end{cases}$$

Moreover to solve the discontinuity of the forces, that can cause some instability in a simulation, we need to consider a linear term proportional to the first derivative of the potential, such as

$$v^*(r) = \begin{cases} v(r) - v(r_c) - \left.\dfrac{dv(r)}{dr}\right|_{r_c} (r - r_c) & r \leqslant r_c \\ 0 & r > r_c \end{cases}$$

Although, the shift methods make the potential quite different from the "true" one and this makes difficult to retrieve the correct thermodynamics proprieties. Thus, even if it can solve some instabilities, it must be carefully used.

Another powerful method is the *switch* method. The general idea is to consider two cut–off radii r_{c1} and r_{c2}. If $r \leqslant r_{c1}$ the "true" forms are used; while for $r > r_{c2}$ the potential is set to zero. For $r_{c1} < r \leqslant r_{c2}$ a *switching function* is considered in order to *smoothly* switch the potential to zero.

It is important to stress out that even the method used to treat the interactions, as the cut–off radii and eventually the switching function, are part of the simulation parameters that are still part of the FF. So they are interdependent with the model parameterizations, and should never be changed without retesting some target properties.

1.4.3 Van der Waals interactions

Van der Waals forces are a set of interactions that are divided into two main contributions: an attracting interaction and a repulsive one. The main contribution to both is due to quantum dynamics effect of the electron cloud interactions through the Pauli exclusion principle and to the instantaneous electrostatic interactions, even if both atoms are neutral, such as dipole–dipole, induced dipole–dipole and induced dipole–induced dipole interactions which, in a more rigorous description should be treated quantum mechanically. Both London dispersion forces, involving polar and non-polar atoms and related to instantaneous multipoles interactions, and hydrogen bonding, due to quantum effects, instantaneous electrostatics interactions and entropy effects, contribute to the attractive part of the potential.

The usual model to treat Van der Waals interactions is a Lennard–Jones potential. The most common exponents for the attractive and repulsive contributions to the potential are 6 and 12, respectively, although 6 and 9 can also be found depending on the system. The general form for a $12-6$ Lennard–Jones potential is the following

$$v(r) = 4\epsilon \left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right) = \frac{C_{12}}{r^{12}} - \frac{C_6}{r^6} \tag{1.4.2}$$

where $C_{12} = 4\epsilon\sigma^{12}$, $C_6 = 4\epsilon\sigma^6$ and r is the pairwise particles distance. ϵ is related to the absolute value of minimum while σ is relented to the position of the minimum of the potential: $r_{min} = 2^{1/6}\sigma$, often referred to by Van der Waals radius.

These constants are proper for each particle pair. The attractive contribution is due to the negative part proportional to r^{-6} while the repulsive one is due to the positive part proportional to r^{-12}. In figure (5) there is a example plot of the function (1.4.2) with $\epsilon = \sigma = 1$.

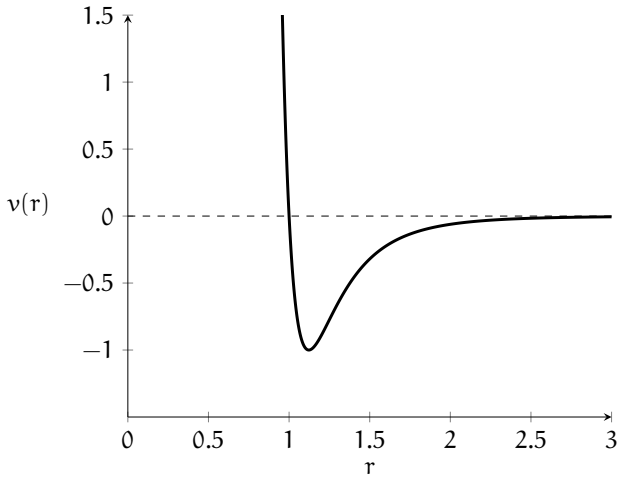


Figure 5: Example plot of a Lennard–Jones function with $\epsilon = \sigma = 1$.

The simplest and computationally most efficient way to treat a Lennard–Jones function, and in general all the short range interactions, is to use the cut–off method together with the shift or switch methods in order to obtain a continuous potential and/or a continuous forces. As we can see from figure (5) the Lennard–Jones potential go to zero rapidly with distance: at $r \sim 2\sigma$ its value is less then 1% of the value in $r \sim \sigma$. A good choice for the cut–off is then of the order of $r_c \sim 2\sigma \div 3\sigma$.

1.4.4 Electrostatic interactions

One of the most important long range interaction is the electrostatic one. Despite the long range characteristic, for purely computational efficient reason, most of the FFs for biomolecular applications treat them in the same way as a short range interaction by a cut–off method[3]. Of course this is an approximation and can lead to a serious issues in those proprieties or systems that strongly depend on the electrostatic interaction. Most issues due to a bad treatment of the Coulomb interactions are related to the development of a good polar solvent model (a good treatment of the electrostatic proprieties of water is really important for biological applications), as well as to the study of the interactions of charged particles with polar solvents, transport processes of charged moieties, calculations of the electrostatic potential inside macromolecules and so on. The loss of computational efficiency in the calculations of the electrostatic energy contribution is that it needs to take into

3 In general, for computational reason, a common choice is to consider the some cut–off for Van der Waals and electrostatic interactions.

account *all particles* in a system, but we can not forget PBC: even all the infinity images of all particles need to be taken into account.

If we consider only the simulation box the energy contribution is

$$U = \frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\frac{1}{4\pi\varepsilon_0}\frac{q_i q_j}{r_{ij}} \tag{1.4.3}$$

where q_i and q_j are the charge of particles i and j, respectively, and r_{ij} is the distance between i and j. But we need also all image boxes. Supposing, for simplicity, that the box is a cube of size L, then we can define a tern of integer numbers $(n_x,\ n_y,\ n_z)$, $n_i = 0, 1, 2, \cdots$ so that the position of all other image boxes, with respect to the central simulation box, is $\vec{n} = L(n_x,\ n_y,\ n_z)$. Then the energy contribution becomes

$$U = \frac{1}{2}{\sum_{n_x,n_y,n_z}^{+\infty}}' \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{4\pi\varepsilon_0}\frac{q_i q_j}{\|\vec{r}_i - \vec{r}_j + \vec{n}\|} \tag{1.4.4}$$

where the prime indicates that for $\vec{n} = 0$, i.e. the energy contribution of the simulation box, we need to exclude the self interaction term: So in inner sum it must be $j \neq i$.

As described above, a cut–off method is a good easy way to solve equation (1.4.3) and sometimes it produces good results. However, the increasing of computer power can lead to develop more rigorous methods to solve equation (1.4.4), even for very large systems. The main problem is that the summation in equation (1.4.4) is *conditionally convergent*[4] and converges extremely slowly so that it would need so many terms to converge that its computational cost would be too high, especially for large systems (of the order of $N \sim 3 \cdot 10^4$). The most important methods developed to solve this problem are based on the *Ewald Summation Method* (ESM). We shall describe those used in this thesis work: the Ewald Summation Method (ESM) itself and the *Particle Mesh Ewald* (PME) method. For a more complete discussion about the advanced methods developed to treat the electrostatic interactions for biological applications the reader is addressed to the Review by Cisneros *et al.* [4].

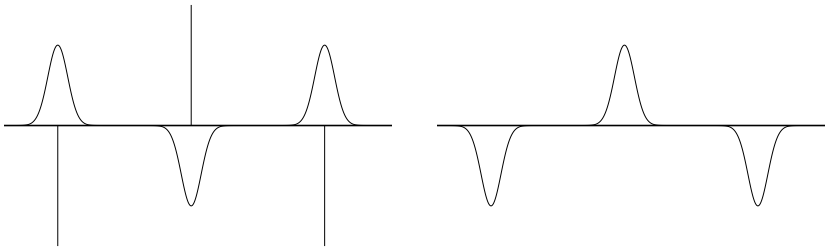
Ewald summation method

The Ewald Summation Method (ESM) is the first method introduced by Ewald for a correct treatment of the electrostatic energy contribution in an ionic crystal that can be modeled as the electrostatic interactions of a periodic charge density. The basic idea is to split the summation in equation (1.4.4) in two series both rapidly convergent. The method is based on the following identity

$$\frac{1}{r} = \frac{f(r)}{r} + \frac{1 - f(r)}{r} \tag{1.4.5}$$

4 A conditionally convergent series contains both positive and negative terms such that the positive or negative term alone form both a divergent series. The sum of a conditionally convergent series depends on the order in which the positive and negative terms are considered.

the trick is to choose a function $f(r)$ that will deal with the rapid variation of the $1/r$ term for small r and the slow decay at long r; in that case the two series can rapidly converge.



(a) Real space (b) Reciprocal space

Figure 6: Schematic illustration of the Ewald Summation Method charge distribution: in (a) point charges (represented by vertical lines) and the neutralizing Gaussian charge distribution; in (b) the counteracts Gaussian distribution.

The ESM for electrostatic interactions works, as illustrated in figure (6), considering each point–like charge in the system surrounded by a neutralizing charge distribution of equal magnitude but opposite sign that decays rapidly to zero. Simplifying the notation for a one dimensional system, the simplest functional form is a Gaussian distribution centered in the position r_i of the point–like charge q_i, of the form

$$\rho_i(r) = \frac{q_i \alpha^3}{\pi^{3/2}} e^{-\alpha^2 (r - r_i)^2} \tag{1.4.6}$$

that obeys the relation

$$\frac{q_i \alpha^3}{\pi^{3/2}} \int_{r_i - \epsilon}^{r_i + \epsilon} e^{-\alpha^2 (r - r_i)^2} \, dr \simeq q_i$$

where $(r_i - \epsilon; r_i + \epsilon)$ is a small interval around r_i. The energy contribution due to this set up, the point–like charge *and* the gaussian charge distribution, is given by

$$U_r = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} {\sum_{n_x, n_y, n_z}}' \frac{q_i q_j}{4\pi\varepsilon_0} \frac{\operatorname{erfc}(\alpha \|\vec{r}_i - \vec{r}_j + \vec{n}\|)}{\|\vec{r}_i - \vec{r}_j + \vec{n}\|} \tag{1.4.7}$$

where $\operatorname{erfc}(x) = 1 - \operatorname{erf}(x)$ is the complementary error function and $\operatorname{erf}(x)$ is the error function. They are given by

$$\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^{+\infty} e^{-t^2} \, dt, \qquad \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \, dt \tag{1.4.8}$$

The point is that the summation involving the complementary error function in equation (1.4.7) is rapidly convergent and it needs very few terms so that a cut–off method can be safely used. The rate of convergence depends on the α parameter,

the bigger is α the more rapidly converge and the shorter can be the cut–off radius. Thus the ESM use the $\mathrm{erfc}(r)$ as $f(r)$ function in equation (1.4.5). Of course since we added a non physical neutralizing charge in the system, in order to restore the real charge distribution, we must consider another distribution, called counteracts charge distribution, of equal magnitude but opposite sign. Considering the identity in equation (1.4.5) this lead to an energy contribution of the form $(1 - f(r))/r$, so, using equation (1.4.8), it is of the form $\mathrm{erf}(r)/r$. Another trick is to compute the former in the *real space*, the latter in the *reciprocal space*, thus considering its Fourier transform. This energy contribution is given by

$$U_f = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k_x,k_y,k_z} \frac{1}{4\pi\varepsilon_0}\frac{4\pi}{L^3k^2}e^{-k^2/(4\alpha^2)}e^{i\vec{k}\cdot(\vec{r}_i-\vec{r}_j)} \tag{1.4.9}$$

where $\vec{k} = 2\pi\vec{n}/L$ are the reciprocal lattice vectors. Even U_f converges rapidly as U_r in equation (1.4.7); then a cut–off method can be safely used. Nevertheless, as opposite to U_r, the smaller is α the shorter can be the cut–off. Clearly a proper *balance* between the real and reciprocal space summation is needed.

Since in equation (1.4.7) even the self interaction with each Gaussian is included we need to add another item for cancel it out; this is done by the self–term

$$U_{self} = -\frac{\alpha}{\sqrt{\pi}}\sum_{i=1}^{N}\frac{q_i}{4\pi\varepsilon_0} \tag{1.4.10}$$

Summarizing, the energy contribution of the electrostatic interactions by the ESM, is computed summing equations (1.4.7),(1.4.9) and (1.4.10) to obtain

$$\begin{aligned} U = \; & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}{\sum_{n_x,n_y,n_z}}' \frac{q_iq_j}{4\pi\varepsilon_0}\frac{\mathrm{erfc}(\alpha\|\vec{r}_i-\vec{r}_j+\vec{n}\|)}{\|\vec{r}_i-\vec{r}_j+\vec{n}\|} + \\ & + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k_x,k_y,k_z} \frac{1}{4\pi\varepsilon_0}\frac{4\pi}{L^3k^2}e^{-k^2/(4\alpha^2)}e^{i\vec{k}\cdot(\vec{r}_i-\vec{r}_j)} + \\ & - \frac{\alpha}{\sqrt{\pi}}\sum_{i=1}^{N}\frac{q_i}{4\pi\varepsilon_0} \end{aligned} \tag{1.4.11}$$

The first line is the real space contribution while the second is the Fourier energy contribution. Since the last self–interaction term is constant it does not affect forces computation. The ESM offers a well defined method to properly treat electrostatic interactions, nevertheless it is quite expensive in term of computational resources. If α and the cut–off are constant, then the computation scales as $\sim N^2$; while if α and the cut–off are dynamically updated it scales as $\sim N^{3/2}$; this can lead to an incompatibility between the Van der Waals interactions cut–off and the electrostatic one thus compromising the efficiency gain. Despite we have described ESM applying it to the electrostatic interactions, it can be used, with some changes, with

all long–range interactions and in general to all energy contributions that decay as r^{-d}, for example, even with the Van der Waals energy contribution.

For biomolecular applications most MD tools set an equal cut-off radius for both Van der Waals interaction and the real part of the Ewald summation (1.4.11) in order to achieve for both a scaling of the order $\sim N$. However in this way the computation of the reciprocal part in the Ewald summation (1.4.11) will be very inefficient as it scales as $\sim N^2$. In order to increase the efficiency of the calculation of the Fourier transform various of advanced methods can be used. They are all based on the use of the Fast Fourier Transform (FFT) method. In this way the reciprocal part can scale as $\sim N \ln N$. Since FFT requires discretized quantities, the idea of such methods, called *Particle Mesh* is to consider the charge density spread on a mesh grid and then evaluate the electrostatic potential via solving the Poisson's equation[5] using fast Poisson solver together with the FFT method; this can be done, for example, exploiting the PBC in order to discretize and make periodic the Poisson's equation Such algorithms include the *particle–particle particle–mesh* method, *Particle mesh Ewald* method, *Fast–Fourier Poisson* method and a recent methodology based on multi–scale mesh grid; the efficiency and accuracy of such mesh–based algorithms depends strongly on the way in which the charges are attributed to mesh points, this makes the methods different. In the following we will describe the one used in this thesis work, the Particle Mesh Ewald (PME) method.

Particle mesh Ewald method

Particle Mesh Ewald (PME) method developed by Darden *et al.* [5] is based on the ESM so that the starting point is equation (1.4.11) in which, as described above, the first part of the Ewald summation is computed in the real space together with the Van der Waals contribution using the same cut–off radius while the reciprocal part is computed using FFT method, in order to have a gain of performance. To do this, first, we need to consider a grid mesh onto which the Gaussian counteracts charge distribution is spread. The basic idea, then, is to calculate the electrostatic energy solving Poisson's equation through FFT methods. The efficiency and accuracy depend on the way the charges are distributed onto the grid. To do this a *charge assignment function*, $W(r)$ is introduced such that, considering for simplicity a one dimensional system, the fraction of a charge at position r assigned to a grid point at position r_p is given by $W(r_p - r)$. Hence, if we have a charge density $\rho(r)$ then the charges at the grid point r_p are given by

$$q_M(r_p) = \int_0^L W(r_p - r)\rho(r)\, dr \tag{1.4.12}$$

where L is the box length and, if h is the grid spacing, $M = L/h$ is the number of mesh point. In figure (7) the charges assignment is schematically represented.

5 Given a charge distribution $\rho(\vec{r})$ then the associated electrostatic potential $\phi(\vec{r})$ can be calculated solving the Poisson's equation $\nabla^2\phi(\vec{r}) = -\frac{1}{\varepsilon_0}\rho(\vec{r})$. If a charge q is at position $\vec{r}$ its electrostatic potential energy is given by $U = q\phi(\vec{r})$.

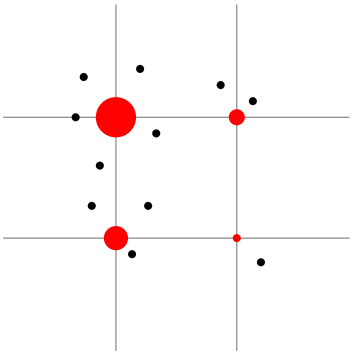


Figure 7: A schematic representation of the charge assignment. The black filled circles are a unit particle charge, while the red ones, are the charges assigned to grid points. The bigger is the circle, the more is the charge.

The assignment function should have the following proprieties: should be an even function and should be normalized in such a way that the sum of the fractional charges equals the total charge of the system. Moreover the best accuracy is obtained with a dense grid in order to reduce as much as possible the discretization of the charge density. However the computational cost increases as the number of grid points! So a balance between efficiency and accuracy is clearly needed.

A nice way to solve the problem of charge assignment is to shift the problem to the discretization of the Fourier transform. This can be viewed as an interpolation problem. Consider the $e^{-i\vec{k}\cdot\vec{r}_j}$ term in the Fourier transform of equation (1.4.9). In general $\vec{r}_j$ does not correspond to a mesh grid point, so that term is not part of a discrete Fourier transform. The idea, thus, is to interpolate it in terms of values of the complex exponential at the mesh points. Switching for simplicity to a one dimensional system, if the mesh grid has $M = L/h$ points, a particle coordinate r_j is located between mesh points $[r_j/h]$ and $[r_j/h] + 1$ where $[]$ denotes the integer part; thus a p–order interpolation of the exponential is of the form

$$e^{-ikr_j} \simeq \sum_{i=1}^{M} W_p\left(\frac{r_j}{h} - i\right) e^{-ikhi}$$

where W_p denotes the interpolation coefficient. A p–order interpolation means that only the p mesh points nearest to r_j contribute to the sum. Assuming a point–like charge distribution the Fourier transform of the charge density is therefore

$$\rho_k \simeq \sum_i e^{-ikhi} \sum_j q_i W_p\left(\frac{r_j}{h} - i\right)$$

we can interpret the above expression as the discrete Fourier transform of the charge density

$$\rho(i) = \sum_j q_i W_p\left(\frac{r_j}{h} - i\right)$$

but using equation (1.4.12), it is nothing that the point–like charge distribution assigned to the mesh point i through the assignment function W_p.

We clearly see that the charge assignment problem is now shifted to the complex exponential interpolation. There are two main methods to make the interpolation:

the *Lagrange interpolation method* and the *Euler SPLINE interpolation method*. The basic idea of the former is to use, as interpolating function, a polynomial function of degree $\leqslant (n-1)$ where n is the number of points to interpolate, that passes through all the n points, and which is constructed with a summation over the *Lagrange basis polynomials* as follow

$$P(x) = \sum_{i=1}^{n} y_i \prod_{\substack{k=1 \\ k\neq i}}^{n} \frac{x - x_k}{x_i - x_k}$$

where $(x_i; y_i)$ are the sets of points to interpolate. The main disadvantage of this method is that, even if $P(x)$ is continuous everywhere, its derivative is not, thus it can lead to some instability in MD simulations. The latter method, that is the most used in MD tools, is based on the concept of *SPLINE interpolation*. Instead of using a unique interpolating function that passes through each point, the SPLINE method uses a *piecewise polynomial function*, called SPLINE, in which each piece is smoothly connected and optimized to interpolate a subset of the points. The Euler SPLINE method use the *exponential Euler SPLINE* that is constructed with the basis of the Euler n–degree polynomials $A_n(x;\lambda)$ generated by the following equation

$$\frac{\lambda - 1}{\lambda - e^z} e^{xz} = \sum_{n=0}^{+\infty} \frac{A_n(x;\lambda)}{n!} z^n$$

where λ is a complex parameter and z is a complex variable. The main properties of such SPLINE is that, it is $n-1$ times analytic, continuously differentiable and then can solve the instability problem of the Lagrange interpolation method. In literature the Euler SPLINE method is referred to as *smooth* Particle Mesh Ewald and the reader is addressed to the article by Essmann *et al.* [6] for more technical details about the interpolation procedure.

Summarizing, the PME method is implemented with the following scheme

- By the interpolation of the complex exponential in the Fourier transform of the Ewald summation, the Gaussian counteracts charge distribution are spread onto the mesh grid and a discretized Fourier transform is obtained;
- Poisson's equation for the discretized charges are solved through the FFT methods;
- The reciprocal energy contribution is obtained considering the inverse Fourier transform;
- Electrostatic forces are computed and assigned to the charged system particles.

The main advantages of the PME algorithm are that the potential energy and forces are smooth functions of the particles positions, offers a good energy conservations, offers a very well balance between accuracy and computational efficiency since it scales as $\sim N \ln N$ and it is easily generalizable to interaction potentials that decay

as r^{-d} such as the Lennard–Jones potential. Nevertheless it does not conserve very well the particles momentum due to Root–Mean–Square (RMS) errors in the forces calculations that have to be cancelled out by removing the forces averages.

1.4.5 Charge representation

Even if some methods, such as the PME one, have been developed to speed up the computation of electrostatic energy contribution one of the main problems of FFs for biomolecular applications which are related to the electrostatic interactions, remains the *charge representation*: The way that the charges of atoms or molecules are assigned to the system particles. The problem arises from the necessity to represent the electrons clouds of atoms and molecules in the system an the interactions which they generate, that is, for instance a purely quantum effect. Nevertheless this is crucial for a better description of most electrostatic phenomena such as polarizability of molecules and polar solvent, solvation shell of charged ions, protein–ligands interaction, ion transport through polar and non–polar medium, self assembly processes and so on.

The mostly used solution is the *atom–centered "partial charge" approximation* in which the full charge density of the molecule is replaced by fractional point–like charges assigned to each atom of the molecule. But now, one has to decide how much of the molecular charge density should be assigned to each atom. Traditionally most FFs assign each atom of a molecule a fixed partial–charge. The most used procedure for extracting partial–charges from molecular wave functions is based on fitting atomic charges with the molecular electrostatic potential, computed with *ab initio* calculation such as *density functional theory*. The fitting procedure consists in minimizing the deviation between the electrostatic potential produced by the assigned charge and the molecular electrostatic potential. Such representation is believed to be an important source of error in the electrostatics treatment. Moreover with fixed charge assignment it is more challenging to take into account those phenomena that involve a transfer of charge inside the molecule, as polarization effect. The use of off–centered charges and/or higher order atomic multipoles can significantly improve the treatment of electrostatics but of course it is necessary a good balance between accuracy and performances since the electrostatic problem can rapidly drive to a loss of efficiency sometimes without a really gain in accuracy.

1.4.6 Polarization

Polarization refers to the redistribution of the electron charge density of a molecule in presence of an external electric field, generated, for example, by charged ions or another molecule. Polarization is responsible for non–addictive attractive inter– or intra–molecular interactions which have many–body characteristics. Induced polarization effects should also be included. These effects have been recognized to have an important role in many biological interactions in which different com-

pounds are present. An increasing number of studies show that the lack of these effects can lead to a serious limitations, particularly, for ionic systems and chemical process that involve different environments such us water and proteins or water and lipids. In MD simulations polarization effects are included using either *implicit* or *explicit* methods.

The implicit method completely avoids the many–body calculation by including a mean polarization effect in the functional form of the interaction potential. The general idea is to surround all the simulation box by a transparent medium with a relative dielectric constant ε_r. In this way the polarization effect is taken into account considering a mean field theory and solving the Poisson's equation to determine the electrostatic potential due to system charges by the substitution $\varepsilon_0 \to \varepsilon_0\varepsilon_r$. Since it avoids many–body calculations, this method gives an incomparable gain in performances bust it must be carefully used. The main disadvantage is that the mean polarization effect is added to all system particles and this wash out all the details about a possible polarization effect in a molecule, for example a protein or a piece of protein. This method can be safely used, for example, when our system is composed principally of one kind of solvent, for example water; but if the simulation box is composed of different chemical environments such as water and lipids or other organic compounds, using the same dielectric constant would lead to seriously incorrect results of the proprieties of the organic component and maybe affect the whole simulation.

The way to correct the above behavior is to use an explicit method. As the name suggests, the polarization effect is taken into account for every molecules in the system by a it proper model included in the FF. The general idea is to add some more internal DOF to a molecule or atom to take into account the movement of charges and/or split the point–like charge assigned, for example, to a chemical group, to a partial charge assigned to each particles of the chemical group itself. This can be done for every molecule or atom in the system and thus it is the optimum to better describe systems with different chemical environments.

1.4.7 Coarse–Grained model

As we have introduced at the beginning of this section, for biomolecular applications two main classes of FFs exist: atomistic FFs and CG FFs. Since the atomistic model takes into account all the atoms in a molecule it is obviously the most real and accurate FF. Nevertheless the number of DOF of the system increases leading to a loss of performances. Moreover, basically, the atomistic FFs are efficient until the physical proprieties can be properly sampled on a time scale of a few microseconds over a length scale of a few nanometers. As the time and length scales increase more and more time is needed to carry out a complete simulation. Unfortunately many biological processes involving lipid membranes and other organic molecules, including synthetic compounds, take place on much longer time and length scales.

One possible solution is to *integrate out* some DOF, preserving those that are relevant for the problem in exam: this procedure is called *coarse–graining*. The

basic units of CG FFs are called *beads*, each representing a group of atoms or a well defined chemical moiety. The size of the group of atoms that is represented by a single bead determines the degree of coarsening of the FF. Even in this case, all the general features described above, apply: functional forms need to be chosen and their parameterization need to be adjusted so as to reproduce the desired target properties. Moreover, in this case, even a *mapping* procedure should be defined as the first step in the development of a CG model: This establishes a link between the atomistic model and the coarse–grain beads. There is not a unique or correct procedure to obtain the mapping because it depends on the desired coarse–graining level, on the time and length scales that one wants to correctly sample and on the properties one wants to reproduce. For biological applications CG FFs are often designed to reproduce specific thermodynamics properties such as surface tension, free energy of partitioning, free energy of hydration and so on, instead of, for example, the structural properties.

In general a CG FF is more computationally efficient than an atomistic one for the following reasons: the DOF of the system are reduced due to the CG procedure and a smaller number of interactions and forces has to be taken into account; bead–bead interactions, which result from the removal of finer structural details, are softer than atom–atom interactions. Thus, vibrational modes are slower, and their sampling can be achieved using larger MD time steps than in atomistic simulations; softer interactions imply a smoother PEF which leads to faster diffusion.

1.5 MARTINI: A COARSE–GRAINED FORCE–FIELD

MARTINI is a CG FF originally developed by Marrink *et al.* [15] for organic solvents and lipids and then extended to proteins [17], carbohydrates [14] and a broad class of polymers [20]. The original aim was to improve the description of the physical and chemical properties of lipid membranes using a CG model. The power of the model was immediately clear and soon its philosophy was changed to develop a FF applicable to a broad range of organic system providing a set of extensively calibrated building blocks to construct a large variety of organic molecules without reparametrizing the FF. This is possible, because, instead of focusing on accurately reproducing structural details of a particular system, the FF is based on accurately reproducing the interaction between polar and non–polar chemical compounds. This is the main target property: The *partitioning free energy* between water and a large number of organic solvents, i.e. the free energy of transfer of chemical moieties from polar and non–polar solvents. These building blocks are representative of the main chemical moieties in an organic system, this is the guide for the mapping procedure.

1.5.1 Mapping

The mapping of the MARTINI beads, is based on a four–to–one scheme that groups four heavy atoms like C, S, O and so on, plus their associated hydrogen atoms,

Table 1: Interaction strength parameter (ϵ). The last one is for the special case $\sigma = 0.62$ nm.

Level	ϵ [kJ/mol]
O	5.6
I	5.0
II	4.5
III	4.0
IV	3.5
V	3.1
VI	2.7
VII	2.3
VIII	2.0
IX	2.0

into a single interaction site. Consistently four water molecules are modeled with one MARTINI bead. An example of the coarse-graining procedure including both atomistic and CG descriptions is shown in figure (8).

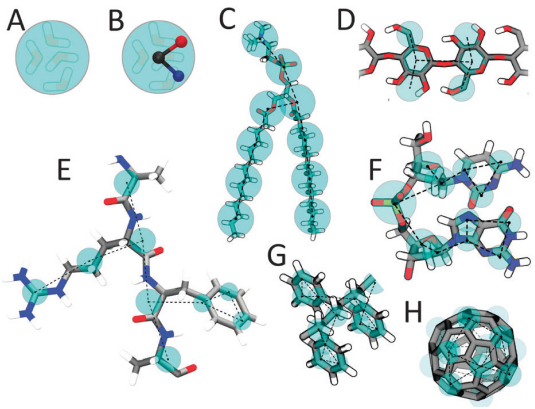


Figure 8: MARTINI mapping and atomistic structures compares of some molecules: (A) Standard water bead, (B) polarizable water bead, (C) DMPC lipid, (D) Polysaccharide fragment, (E) Peptide, (F) DNA fragment, (G) Polystyrene fragment and (H) Fullerene molecule. In all cases MARTINI CG beads are shown as cyan transparent beads overlaying the atomistic structure. Taken from [16].

There are four main bead types: polar (P), non–polar (N), apolar (A) and charged (Q). Each bead type has a number of subtypes to take into account a more accurate representation of the chemical nature of the moieties due to underlying the specific atomistic structure. These subtypes are distinguished by the hydrogen bonding capabilities: donor (d), acceptor (a), both donor and acceptor (da) and none (0) and/or by their degree of polarity: lowest polarity (1), $\cdots$, highest polarity (5).

1.5.2 Interactions potential

VAN DER WAALS INTERACTIONS The functional form describing pairwise Van der Waals interaction is a Lennard–Jones $12-6$ potential as in equation (1.4.2). For most beads the σ parameter is set equal to 0.47 nm except for the Q–C_1 and Q–C_2 interactions for which $\sigma = 0.62$ nm. This is consistent with reproducing the hydration shell when a charged bead (Q) is dragged into an apolar medium. The strength of the interactions is instead dived into ten levels, reported in table (1). The association of the interactions strength with the MARTINI beads is shown in figure (9).

		Q				P					N				C				
	sub	da	d	a	0	5	4	3	2	1	da	d	a	0	5	4	3	2	1
Q	da	O	O	O	II	O	O	O	I	I	I	I	I	IV	V	VI	VII	IX	IX
	d	O	I	O	II	O	O	O	I	I	I	III	I	IV	V	VI	VII	IX	IX
	a	O	O	I	II	O	O	O	I	I	I	I	III	IV	V	VI	VII	IX	IX
	0	II	II	II	IV	I	O	I	II	III	III	III	III	IV	V	VI	VII	IX	IX
P	5	O	O	O	I	O	O	O	O	O	I	I	I	IV	V	VI	VI	VII	VIII
	4	O	O	O	O	O	I	I	II	II	III	III	III	IV	V	VI	VI	VII	VIII
	3	O	O	O	I	O	I	I	II	II	II	II	II	IV	IV	V	V	VI	VII
	2	I	I	I	II	O	II	II	II	II	II	II	II	III	IV	IV	V	VI	VII
	1	I	I	I	III	O	II	II	II	II	II	II	II	III	IV	IV	IV	V	VI
N	da	I	I	I	III	I	III	II	II	II	II	II	II	IV	IV	V	VI	VI	VI
	d	I	III	I	III	I	III	II	II	II	II	III	II	IV	IV	V	VI	VI	VI
	a	I	I	III	III	I	III	II	II	II	II	II	III	IV	IV	V	VI	VI	VI
	0	IV	IV	IV	IV	IV	IV	IV	III	III	IV	IV	IV	IV	IV	IV	IV	V	VI
C	5	V	V	V	V	V	V	IV	IV	IV	IV	IV	IV	IV	IV	IV	IV	V	V
	4	VI	VI	VI	VI	VI	VI	V	IV	IV	V	V	V	IV	IV	IV	IV	V	V
	3	VII	VII	VII	VII	VI	VI	V	V	IV	VI	VI	VI	IV	IV	IV	IV	IV	IV
	2	IX	IX	IX	IX	VII	VII	VI	VI	V	VI	VI	VI	V	V	V	IV	IV	IV
	1	IX	IX	IX	IX	VIII	VIII	VII	VII	VI	VI	VI	VI	VI	V	V	IV	IV	IV

Figure 9: Interaction strength association matrix for the MARTINI bead types and subtypes. Taken from [15].

ELECTROSTATIC INTERACTIONS Electrostatic charges are assigned using the atom–centered approximation, as described in 1.4.5. In this case, charges are no more fractional but are empirically assigned at the center of the beads trying to follow as much as possible the net charge of the associated chemical moieties. A special case is water, modeled as a P_4 bead, since it interacts only with Van der Waals interaction so that the polarizability effect is not very well described. To fill this lack it is used an implicit medium with a dielectric constant $\varepsilon_r = 15$. However, as we will see later in 1.5.5, to avoid the problems with the implicit medium described in 1.4.6, especially for lipid membranes for which the dielectric constant in the hydrophobic region is much smaller, Yesylevskyy *et al.* [24] have developed a more sophisticated CG water model, called Polarizable Water (PW), to take into account a better water behavior.

BONDED INTERACTIONS They includes only bond length and an angle harmonic contributions. The former is modeled with a harmonic potential as the first term in equation (1.4.1), with the same bond constant for all bead types: $k^b =$

1250 kJ/(mol nm^2) and an equilibrium distance of $l_0 = 0.47$ nm. The later is modeled as a cosine–type harmonic potential

$$U = \frac{1}{2}k^a(\cos(\theta) - \cos(\theta_0))^2$$

whose parameters are: $k^a = 25$ kJ/mol and $\theta_0 = 180°$ for aliphatic chains; $k^a = 45$ kJ/mol and $\theta_0 = 120°$ for `cis` double bonds and $k^a = 45$ kJ/mol and $\theta_0 = 180°$ for `trans` unsaturated bonds. Moreover, especially for ring systems, an improper dihedral angle harmonic potential can be used to prevent out of plane distortion. The form is

$$U = k_{id}(\theta_{ijkl} - \theta_0)^2$$

where θ_{ijkl} denotes the angle between the planes described by atoms i, j, k and j, k, l; k_{id} and θ_0 are, as usual, the force constant and equilibrium angle.

1.5.3 Simulation parameters

MARTINI FF was originally developed using a shifted cut–off scheme for both Lennard–Jones and electrostatic potentials with a cut–off radius $r_c = 1.2$ nm. The Lennard–Jones potential was shifted from $r_s = 0.9$ nm to r_c while from $r_s = 0.0$ nm to r_c for the electrostatic potential. The neighbor list is constructed as described in the first part of 1.3.4 with a refresh rate of 10 MD steps. Recently the more efficient Verlet cut–off scheme was tested by Marrink *et al.* [16] and used with the MARTINI FF with a cut–off radius of $r_c = 1.1$ nm, a Verlet buffer tolerance of 0.005 kJ/(mol·ps) and a minimum refresh rate of 10 MD steps (often, depending on the hardware, it can be dynamically increased to 30 or 40 MD steps getting better performances). Moreover, the treatment of the electrostatic interaction can be safely updated to the PME method together with the Verlet cut–off scheme. This new set–up was largely tested by Yesylevskyy *et al.* [24]. In this case the cut–off radius was set to $r_c = 1.2$ nm with the same Verlet buffer tolerance; the PME grid spacing was set to have a lower bound of 0.12 nm and the interpolation was set to a fourth-order. Moreover, with the use of PW, as we shall see, the dielectric constant should be reduced to $\varepsilon_r = 2.5$. In all cases a time step up to 40 fs is suitable for a great number of applications, but 20 fs is the most powerful choice in terms of performance and accuracy balancing. It should be clear that changing these simulation parameters must be followed by a retest of the main properties of the MARTINI FF.

1.5.4 Parametrization

In order to parametrize the MARTINI CG FF a set of thermodynamics properties, obtained from MD simulations are compared and fitted against those experimentally measured. These properties are the *free energies of vaporization*, *hydration* and *partitioning* between water and a set of organic compounds such as hexadecane (H), chloroform (C), ether (E) and octanol (O). The free energy of hydration was

obtained from the partitioning of CG compounds between bulk water in equilibrium with its vapor. Similarly the free energy of vaporization was obtained considering a simulation box with the selected CG compounds in equilibrium with their vapor. From the equilibrium densities of the particles in both the phases the related free energies can be computed from

$$\Delta G = k_B T \ln \left(\frac{\rho_{vap}}{\rho_{bulk}} \right)$$

All the simulations were performed in a canonical NVT ensemble.

Instead, the partitioning free energy between water and an organic solvent was obtained in a NPT ensemble, considering a simulation box half filled of water and half of the organic solvent. Then a small fraction of the CG particle for which the partitioning free energy is to be computed, was placed in the simulation box. From the equilibrium densities of the particles in water ρ_{wat} and in organic solvent ρ_{oil}, the free energy of transfer can be computed from

$$\Delta G^{part}_{SW} = k_B T \ln \left(\frac{\rho_{oil}}{\rho_{wat}} \right)$$

where S indicate the organic solvent.

In figure (10) a summary of the results is reported. As one can see the model has bad performances for what concerns the free energies of vaporization and hydration, which are too high with respect to experimental data. Instead, the partitioning free energies match very well. Thus the model is not very accurate for vapor–liquid systems, but as long as one does not study those systems, the partitioning free energy is much more important than the other free energy contributions.

1.5.5 Polarizable Water model

Water play a crucial role in any biomolecular systems thus it is important to correctly describe its behavior. Since the MARTINI water model does not directly take into account the electrostatic interaction between water and the other molecules because it does not have any charge and it interact only via Van der Waals interaction, thus a simple implicit medium is used to take into account the main effects of water, screening and polarizability. However any biomolecular process involve charged species moving between regions of different dielectric constant. Due to the change in electrostatic screening between those environments, the strength of the interaction between the moving charges and the surrounding molecules also changes, but this effect can not be consider in an implicit medium model. Thus can have important consequences for the way biological activity is controlled. In order to capture the inhomogeneous nature of the dielectric response an explicit medium has to be used.

In the same fashion of the MARTINI philosophy, Yesylevskyy *et al.* [24] have developed a Polarizable Water (PW) model that better describe the real behavior

			ΔG^{vap}		ΔG^{hyd}		ΔG^{part}_{HW}		ΔG^{part}_{CW}		ΔG^{part}_{EW}		ΔG^{part}_{OW}	
type	building block	examples	exp	CG	exp	CG	exp	CG	exp	CG	exp	CG	exp	CG
Q_{da}	H_3N^+—C_2—OH	ethanolamine (protonated)				−25		< −30		−18		−13		−18
Q_d	H_3N^+—C_3	1-propylamine (protonated)				−25		< −30		−18		−13		−18
	NA^+OH	sodium (hydrated)				−25		< −30		−18		−13		−18
Q_a	PO_4^-	phosphate				−25		< −30		−18		−13		−18
	CL^-HO	chloride (hydrated)				−25		< −30		−18		−13		−18
Q_0	C_3N^+	choline				−25		< −30		−18		−13		−18
P_5	H_2N—C_2=O	acetamide	sol	sol	−40	−25	−27	−28	(−20)	−18	−15	−13	−8	−10
P_4	HOH (× 4)	water	−27	−18	−27	−18	−25	−23		−14	−10	−7	−8	−9
	HO—C_2—OH	ethanediol	−35	−18	−33	−18	−21	−23		−14		−7	−8	−9
P_3	HO—C_2=O	acetic acid	−31	−18	−29	−18	−19	−21	−9	−10	−2	−6	−1	−7
	C—NH—C=O	methylformamide	−35	−18		−18		−21		−10		−6	−5	−7
P_2	C_2—OH	ethanol	−22	−16	−21	−14	−13	−17	−5	−2	−3	1	−2	−2
P_1	C_3—OH	1-propanol	−23	−16	−21	−14	−9	−11	−2	−2	0	1	1	−1
		2-propanol	−22	−16	−20	−14	−10	−11	−2	−2	−1	1	0	−1
N_{da}	C_4—OH	1-butanol	−25	−16	−20	−9	−5	−7	2	0	4	2	4	3
N_d	H_2N—C_3	1-propylamine	−17	−13	−18	−9	(−6)	−7	(1)	0	(−3)	2	(3)	3
N_a	C_3=O	2-propanone	−17	−13	−16	−9	−6	−7	1	0	−1	2	−1	3
	C—NO_2	nitromethane	−23	−13	−17	−9	−6	−7		0		2	−2	3
	C_3=N	proprionitrile	−22	−13	−17	−9	−5	−7		0		2	1	3
	C—O—C=O	methylformate	−16	−13	−12	−9	(−6)	−7	(4)	0	(−1)	2	(0)	3
	C_2HC=O	propanal		−13	−15	−9	−4	−7		0	2	2	3	3
N_0	C—O—C_2	methoxyethane	−13	−10	(−8)	−2	(1)	−2		6	(3)	6	(3)	5
C_5	C_3—SH	1-propanethiol	−17	−10		1		5		10		10		6
	C—S—C_2	methyl ethyl sulfide	−17	−10	−6	1	(7)	5		10		10	(9)	6
C_4	C_2=C_2	2-butyne	−15	−10	−1	5		9		13		13	9	9
	C=C—C=C	1,3-butadiene		−10	2	5	11	9		13		13	11	9
	C—X_4	chloroform	−18	−10	−4	5	(7)	9	14	13		13	11	9
C_3	C_2=C_2	2-butene		−10		5		13		13		13	13	14
	C_3—X	1-chloropropane	−16	−10	−1	5	12	13		13		13	12	14
		2-bromopropane	−16	−10	−2	5		13		13		13	12	14
C_2	C_3	propane	gas	−10	8	10		16		15		14	14	16
C_1	C_4	butane	−11[b]	−10	9	14	18	18		18		14	16	17
		isopropane	gas	−10	10	14		18		18		14	16	17

Figure 10: Results summary: free energies of vaporization ΔG^{vap}, hydration ΔG^{hydr} and partitioning ΔG^{part} between water (W) and organic solvents (hexadecane (H), chloroform (C), ether (E) and octanol(O)) compared to experimental values. Experimental properties in parentheses are estimates obtained from comparison to similar compounds. The statistical accuracy of the free energies obtained from the simulations is ±1 kJ/mol. [b] The temperature for the experimental data is 273 K. Taken from [15].

(standard model) (PW model)

Figure 11: Schematic representation of the PW bead. Shaded orange spheres correspond to the Van der Walls radii of the central neutral particle *W*. The blue particle is the positively charged while the red is the negatively charged.

of water. As before four water molecules is associated to one PW bead. The new water bead actually consists of three particles instead of one in the standard MARTINI model. In figure (11) the topology of the PW and a comparison between the old model is shown. The central particle *W* is neutral and interacts with other particles in the system with only Lennard–Jones potential, just like the standard water

bead thus as a P_4 MARTINI bead (see figure (9) for the interactions matrix). There are two additional particles, namely *WP* and *WM*, that are bound to the central particle and carry a positive and negative charge $|q| = 0.46e$ respectively, where $e = 1.60217653(14) \cdot 10^{-19}$ C is the unit electron charge. They interact with other particles in the system by the Coulomb interaction only. The bonds $W - WP$ and $W - WM$ are constrained to have a fixed distance $l = 0.14$ nm. The electrostatic interaction between *WP* and *WM* inside the same bead are exclude thus they are transparent toward each other and they can rotate around the *W* particle. As a consequence the dipole momentum of the water bead depends on the relative angular position θ of *WP* and *WM*: It can vary from zero ($\theta = 0$), to $2ql$ ($\theta = \pi$). A harmonic angle potential with equilibrium angle fixed to $\theta_0 = 0$ and a force constant $k_\theta = 4.2$ kJ/(mol·rad^2) is added to control the rotation of *WP* and *WM* particles around the *W* particle so to adjust the distribution of the dipole momentum. The value of the equilibrium angle is consistent with the fact that in an apolar medium the total dipole momentum of a water molecule is zero. Since in this model the screening and polarization effects are treated explicitly the global dielectric constant is then reduced from $\varepsilon_r = 15$, used in the standard MARTINI , to $\varepsilon_r = 2.5$. Moreover, since the PW beads attract each other stronger then the standard water beads because of additional electrostatic interactions the strength ϵ_{WW} of the Lennard–Jones interaction between *W* particles must be reduced. They found that change the Lennard–Jones strength from an I level to an III level can do the trick (see table (1) for the various interaction levels). While σ remains set to 0.47 nm.

The parametrization of q, k_θ and ϵ_{WW} are obtained, in addition to the basic target properties of the MARTINI FF, also trying to reproduce the dielectric constant, density and dipole momentum of a pure water phase. For a more details about the parameterizations and testing methods the reader is addressed to the article by Yesylevskyy *et al.*[24]. An important question is the use of the PME method to treat the long range electrostatic interaction. The authors found that, despite the loss of performance, in addition to the PW model, the use of the PME method contributes to a more realistic description of the processes involved in a biomolecular environments. In particular, some interesting results of utility for this thesis work, concern a better description of some properties of lipid membranes. One is the translocation of charged ions through a lipid bilayer that is described in a more realist detail: The authors found that the simulations with PW and PME are approaching the atomistic results better then the standard MARTINI FF. Another important phenomenon about lipid membranes consist in the electroporation of membrane by PW beads due to an electric field across the membrane (for example, created by a cross membrane ions imbalance) and, relented to this, the translocation of ions across the membrane helped by a so called "*water finger*" a water defect inside the membrane that seems to be the trigger of the ions translocation. Such phenomenons are clearly evident in atomistic simulation however never occurred using the standard MARTINI FF thus the importance to use the PW model together with PME method.

1.5.6 Limitations of MARTINI FF

As we have seen in 1.4.7 the CG FFs are computationally advantages, however a price has to be paid. Although the MARTINI FF is still a finer CG FF, some limitations are shared with other CG models at a fundamental level, such as the chemical and spatial resolution, which are both limited compared to atomistic models. One of the most important question is the DOF reduction: This affects the entropy of the system which is underestimated and, as consequence of the MARTINI parameterizations, also the balance between enthalpy is not correct described. Since the MARTINI model is built with the constraint that the partitioning free energy $\Delta G = \Delta H - T\Delta S$ must be conserved with the atomistic case, together with an intrinsic loss of entropy, the enthalpic contribution must decrease with respect to the atomistic model[6]. Moreover this mean that the temperature dependence of a CG model is a priori not correct, anyway this is not our case since we use only a NVT or a NPT ensemble with an external thermostat coupling.

Another consequence of CG FFs is related to the PEF that becomes smoother respect to the atomistic case. This effectively results in more sampling of the energy landscape in a given time period, speeding up the kinetics of the system and allowing the use of an higher time steps with longer simulation times. However the speed–up is not easily predictable and is not likely to be the same for different systems and maybe it is dependent on the type of molecule. Nevertheless for the MARTINI FF an average scaling factor of 4, based on lateral diffusion coefficients of lipids in membranes, is commonly used, of course with some care. Another smaller source of errors is due to the choice of masses: Since ensemble properties is not affected by particles mass, in order to increase the efficiency, all the MARTINI beads have the same mass of 72 amu. This leads in some uncertainty in the dynamics of the system making the time scaling for different beads non–trivial.

A problem involving the Lennard–Jones potential as a model of Van der Waals interaction in MARTINI , is that the steep repulsion leads to an over–structuring of fluids compared to atomistic models. The direct and most evident implication is the melting point of the water that is 290 ± 5 K. A practical partial solution is the use of the so called "anti–freeze" particles named BP_4 type. The Lennard–Jones interaction between these particles and water is modified with a slightly larger Van der Waals radius parameter, $\sigma = 0.57$ nm and a stronger interaction to be a level O (see table (1) for the interaction level). Marrink *et al.* suggest that a mole fraction of $n_{af} = 0.1$ is sufficient to prevent freezing without affecting the other properties of water. Some other properties of water are not accurate described such us the surface tension of air/water interface that leads in problem to water/oil interfaces formation. Using the PW model these properties improve slightly. For a more comprehensive discussion about the limitations of the MARTINI FF the reader is addressed to the review by Marrink *et al.* [16].

6 If a NVT ensemble is used the correct potential is $\Delta A = \Delta U - T\Delta S$, thus the incorrect balance is between the internal energy and the entropy.

1.6 ADVANCED SAMPLING METHODS

The quantity of particular importance for the equilibrium statistical mechanics of which we are interested to obtain with MD technique is the free energy function: the Gibbs free energy G for the isobaric–isothermal ensemble and A the Helmholtz free energy for the canonical ensemble. Being related to the partition function of an ensemble, the free energy is the generator through which other thermodynamic quantities are obtained via differentiation. However often we are particularly interested in the free energy difference between two thermodynamic states, instead of the absolute value. This because free energy differences are the driving force of any process and they tell us, for example, if a chemical reaction occurs spontaneously, whether a given solute is hydrophobic or hydrophilic, if a protein conformational change take place or whether some molecules in water solution are able to self–assembles into a more complex system and so forth. Moreover, often, we are interested also in the Free Energy Surface (FES): the free energy in function of some generalized coordinates, called Collective Variables (CVs) of the system. These small set of variables can describe, in a simple and useful manner, some chemical, thermodynamic or mechanical processes that take place in the system, for example, the free energy in function of the distance between the COM of two molecules give us information about their attraction or repulsion and if they form a bound state; or in function of a rotational angle of a molecule bond to obtain all the possible stable configurations and so on. Thus the FES provides a map of the stable conformations, the relative stability of these conformations and the barrier heights that must be crossed for the processes to take place. It is necessary to stress out that in many cases one do not know exactly *a priori* the FES about a certain process and so one want to know it also for understand if other minima energy configurations exist, in addition to those known, how stable they are and what are the energy barriers to go from one minima state to an other[7]. Despite this, the calculation of free energy difference of two thermodynamic states (that leads an *a priori* knowledge of the two stable states) and the calculation of the FES are one of the main challenges in MD simulations for biomolecular applications.

Lets us suppose that we are interested in the FES of the CV $s(\vec{r})$ and we are working in a isobaric–isothermal ensemble with an isotropic system. Following section 1.2, the Gibbs free energy along the CV is obtained as

$$G(s) = -k_B T \ln Q(s) \tag{1.6.1}$$

where $Q(s)$ is the partition function with all other DOF expect for $s(\vec{r})$, integrated out. This leads as follow

$$Q(s) = \frac{1}{\mathcal{Z}_{NpT}} \int_0^{+\infty} dV \int_\Omega e^{-\beta(\mathcal{H}(\vec{x})+pV)} \delta(s(\vec{r}) - S)\, dx$$

7 Sometimes thought the cinematic of an MD simulation one can gamble an estimate of the functional form of the FES, looking for the relative probability to stay in a meta–stable state rather then the other. Clearly this means that both meta–stable states are to be sampled, as we shall see later it depends of the height of the energy barriers.

since $s(\vec{r})$ do not depends on particles momenta, from equations (1.1.10) and (1.2.6), it can be rewritten as

$$Q(s) = \frac{\int_\Omega e^{-\beta U(\vec{r})} \delta(s(\vec{r}) - s)\ d\mathbf{r}}{\int_\Omega e^{-\beta U(\vec{r})}\ d\mathbf{r}} \tag{1.6.2}$$

where $U(\vec{r})$ is the PEF. $Q(s)ds$ can be interpreted as the probability of finding the system with $s(\vec{x})$ between s and $s + ds$. Since this equation contains a direct phase space integration can be rewritten in a more useful manner using the ensemble averages then, using the ergodic theorem in equation (1.2.1), as a time averages:

$$Q(s) = \langle \delta(s(\vec{r}) - s) \rangle = \lim_{t \to +\infty} \frac{1}{\tau} \int_0^\tau \delta(s(\vec{r}(t)) - s)\ dt \tag{1.6.3}$$

MD simulations give us the possibility to sample a given ensemble via computational methods, in principle, in order to calculate the integrals in equation (1.6.3). Unfortunately, since the time can not be infinity, the main problem related to MD simulations is whether we are able to correctly sample *all* the phase space of an ensemble in order to compute the ensemble average. Clearly the answer depends on the system in exam, if it is really simple maybe we can do that, otherwise probably no or it takes too time and/or we are not able to collect sufficiently data. This sampling problem can be summarized as follow: Regions in phase space around a local minimum of the PEF (then a local minimum even in the FES) are typically sampled well, whereas regions of higher energy are sampled rarely. Despite they leads with a small contribution to the partition function, due to Boltzmann factor, a bigger challenge, related to overcome regions of higher energy, is to sample other minima of the PEF, that can be of lower or same energy and instead can leads to a important contribution to the ensemble averages. This is the *rare events problem*. When the system is moving in the PEF landscape the only way to escape from a local minimum is due to thermal fluctuations and so energy barriers that are higher then $\sim k_B T$ have a small probability of being overcame. Then the sampling problem! Moreover the energy landscape, even for a small molecule, it is extremely wrinkled and the large number of free energy minima is far more than can be sampled in a typical MD simulation. Thus the principal reason for the use of the just introduced CVs, they can limit the sampling necessity to those regions of phase space that are most important to the process under study, hoping that, as the DOF reduction in a CG FF, the limited sampled regions are sufficient to correctly describe the process.

Several methods are developed and are still in development in order to solve the just described sampling problem. This method are all based on advanced sampling techniques that allow us to

- Escape from a local energy minima, in order to explore other regions of the phase space;
- Calculate free energy difference between two thermodynamic states;
- Compute the FES along one or a small set of CVs;

- Try to capture and study the transition state of a process (in function of some CVs).

The common basic idea is to introduce additional DOF along the CVs in order to modify or add some potential energy, to the PEF that drive the system from one state to another. This additional potential is called *biased potential* and those advanced sampling methods are grouped in the so called *biased* MD to distinguish from the *un–biased* MD in which the system is not driven. In particular we describe only those methods used in this thesis work: *Umbrella sampling method* and *Metadynamics*. For a more comprehensive discussion the reader is addressed to the review by Kästner [9] for the former and to the review by Laio and Gervasio [10] for the latter. Moreover in the books of Tuckerman [22] other methods are described.

1.6.1 Umbrella sampling

Umbrella sampling method was developed by Torrie and Valleau and today is one of the most mature and broadly accepted method for calculating free energy differences. The basic idea is to drive the system from a known state A to an other known state B through a deterministic path defined on the CV $s(\vec{r})$ chosen to describe the process. This methods is well suitable for one CV otherwise the computational performance degrades rapidly with the number of CVs. The idea is to divide the path into a discrete number of segment N_w and take a subset $\{s_i\}$ of the values assumed by the continuos CV from state A to B along the path. For each of those target values s_i a bias potential $w_i(s)$ depending only on s is added to PEF in order to restrain the system to the target value. Then MD simulations are performed for each value. Those N_w independent simulations are commonly called *windows* for which the CV assume different values. Then, with an analysis method, all data collected with the windows are combined together to compute the FES along the chosen CV. In figure (12) an example of the windows selection is shown.

Suppose $w_i(s)$ to be the bias potential applied on window i in function of the CV $s(\vec{r})$. Then the biased PEF of window i become $U_i^b(\vec{r}) = U(\vec{r}) + w_i(s)$; where the superscript b denotes biased quantities. With the substitution $U(\vec{r}) \to U_i^b(\vec{r})$ in equation (1.6.2) the corresponding biased partition function integrated over all DOF but s and exploiting the Dirac delta properties, yields to

$$Q_i^b(s) = e^{\beta w_i(s)} \frac{\int_\Omega e^{-\beta U(\vec{r})} \delta(s(\vec{r}) - s)\ d\mathbf{r}}{\int_\Omega e^{-\beta (U(\vec{r}) + w_i(s(\vec{r})))}\ d\mathbf{r}}$$

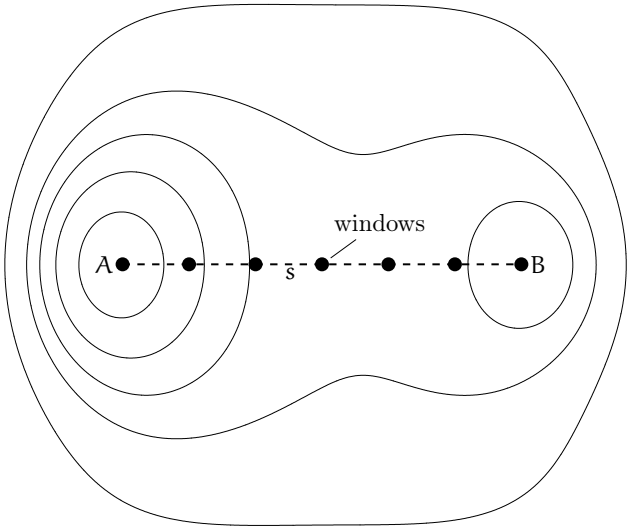


Figure 12: Example of an umbrella sampling windows selection. The contour plot represent a 2–minima free energy profile through which the system is evolved. The dashed line is the path along the CV $s(\vec{r})$ that connect state A to B. The black filled circles represent the selected windows for different values of s.

In order to use equation (1.6.1) to obtain the FES, we need to recover the un–biased partition faction $Q_i(s)$ like in equation (1.6.2). This can be done (see [9]), obtaining the following expression

$$\begin{aligned} Q_i(s) &= Q_i^b(s) e^{\beta w_i(s)} \frac{\int_\Omega e^{-\beta U(\vec{r})} e^{-\beta w_i(s(\vec{r}))} \, dr}{\int_\Omega e^{-\beta U(\vec{r})} \, dr} = \\ &= Q_i^b(s) e^{\beta w_i(s)} \left\langle e^{-\beta w_i(s)} \right\rangle \end{aligned} \tag{1.6.4}$$

then the FES for window i is obtained simply by equation (1.6.1)

$$G_i(s) = -k_B T \ln Q_i^b(s) - w_i(s) + C_i$$

where $Q_i^b(s)$ is obtain as an ensemble average via the biased MD simulation like equation (1.6.3) and $w_i(s)$ is a known function. C_i is an addictive constant independent of s that connect the free energy curves $G_i(s)$ of different windows. As we shall see, in order to combine more windows to calculate a global FES, all set $\{C_i\}$ must be computed.

WEIGHTED HISTOGRAM ANALYSIS METHOD Different analysis methods have been developed in order combine informations collected from the N_w windows. The aim of those methods is to compute the global un–biased partition function $Q(s)$ from the set of biased partition function $\{Q_i(s)\}$ in order to extract the global FES $G(s)$. The most commonly used method is the Weighted Histogram Analysis Method (WHAM) [21], [8]. The philosophy behind is to minimize the statistical error in the calculation of $Q(s)$. For each of the N_w biased windows a set of his-

tograms $\{h_i(s)\}$ with n_i bins, representing the set $\{Q_i(s)\}$, is recorded. The global distribution function is computed as follow

$$Q(s) = \sum_{i=0}^{N_w} p_i h_i(s), \qquad \sum_{i=0}^{N_w} p_i = 1 \tag{1.6.5}$$

where p_i are weights chosen to minimize statistical error on $Q(s)$. Those leads to

$$p_i = \frac{n_i e^{-\beta(w_i(s)-C_i)}}{\sum_{j=0}^{N_w} n_j e^{-\beta(w_j(s)-C_j)}} \tag{1.6.6}$$

where n_i is the total number of independent bins in the $i-$th histogram. The problem now is to compute the set of constants $\{C_i\}$. We can not use the integrals in equation (1.6.4) for the problems already described at the begging of this section, but the second line give us an idea. The ensemble average can be computed using the global distribution function $Q(s)$ as follow

$$e^{-\beta C_i} = \left\langle e^{-\beta w_i(s)} \right\rangle = \int Q(s) e^{-\beta w_i(s)} \, ds \tag{1.6.7}$$

Because $Q(s)$ in equation (1.6.5) depends on the set of constants $\{C_i\}$ and *vice versa*, both equations must be solved in a iterative self–consistent manner. A first guess of set $\{C_i\}$ is used to compute the weights from equation (1.6.6) then from equation (1.6.5) $Q(s)$ is computed and it is used to obtain a new set of constants $\{C_i\}$ from equation (1.6.7) and so on until both equations are satisfied. When the iteration procedure is completed the global FES $G(s)$ is obtained from equation (1.6.1). One important consideration about WHAM is that each window must be sufficient overlap between the distribution of the chosen value of the CV otherwise the statistical error due to the combining procedure can be too high or the iterative producer itself can lead to convergence problem. In general a set of windows are performed then if they are not sufficient other windows are carry out.

BIAS POTENTIAL The bias potential is chosen such that sampling along the CV is uniform. Obviously $w(s) = -G(s)$ is the best optimal choice, unfortunately $G(s)$ is not *a priori* known. Together with WHAM a simple harmonic bias potential is most commonly used for its simplicity. In order to restrain the system to the target value s_i of the CV $s(\vec{r})$ along the path chosen to connect state A and B, each window is biased with a harmonic potential of the form

$$w_i(s) = \frac{1}{2} K (s - s_i)^2$$

The choice of K, the strength of the bias potential, is a critical point. K has to be large enough to drive the system over the barrier. However too large K cause too narrow distribution that can cause overlap problems, thus the necessity for K to be as small as possible to allow for much overlap between windows.

The implementation of the umbrella sampling method can be summarized in the following procedure

- A CV that well describe the transition process from state A to B and a connecting path are chosen;
- A subset of the values assumed by the CV along the path are taken and for each a biased MD simulation is performed (windows)[8];
- Using WHAM the set of biased partition functions $\{Q_i(s)\}$ are combined in order to compute the global partition function $Q(s)$. Then $G(s)$ is calculated;
- If the biased partition functions are not well overlapping then more windows have to be performed.

1.6.2 Metadynamics

The metadynamics method was originally developed by Parrinello and Laio [11] with the first aim to accelerate the escaping from a free energy minima in order to explore other phase space regions. Soon the success of the methods leads a creation of an unified framework for computing free energies in function of few CVs and accelerating rare events. The main advantages respect to umbrella sampling method is that several CVs, instead of only one, can be simultaneously used without affecting the simulation performance. The basic idea of the metadynamics is to enhance the dynamics of a system along some CVs simply by filling the corresponding energy minimum which an history–dependent bias potential, in order to samples a larger and larger portion of the phase space. If the deposited energy is sufficient the system is favorably disposed to overcome the energy barrier to go to another meta–stable state. This novel idea is moreover supported by the assumption of Parrinello and Laio, based on experimental and heuristic results, that iteratively summing the deposited potential during the biased MD simulation leads to an estimator of the FES along the chosen CVs in the region explored. If $\vec{s}(\vec{r}) = (s_1(\vec{r}), \cdots, s_n(\vec{r}))$ is the CVs vector, where n is a small number, and $w(\vec{s}, t)$ is the bias potential deposited every τ MD time steps the "metadynamics" history–dependent potential acting on the system at a time t is given by

$$w_M(\vec{s}, t) = \sum_{\substack{i=0 \\ i\tau<t}} w(\vec{s}, i\tau)$$

where t is the time in unit of MD time step. The time dependence in the bias potential $w(\vec{s}, t)$ is needed since it have to depend on the values assumed by the CVs at some previous MD steps. The Parrinello and Laio assumption yields to the following expression

$$\lim_{t\to+\infty} w_M(\vec{s}, t) \simeq -G(\vec{s}) + C \tag{1.6.8}$$

8 Since each biased simulation is independent they can be performed in parallel with the other.

where C is an addictive constant. Since the history–dependent potential iteratively compensates the underlying FES, the system evolved with metadynamics *tends to escape from any energy minima via the lowest saddle point*. Thus, to the contrary of umbrella sampling metadynamics is suitable, not only to compute efficiently FES, but also to explore new reaction path and accelerate the observation of rare events. If the CVs are chosen sensibly the system will quickly find its way over the lowest free energy saddle point and evolve over the next minimum as it would do in a *very long* MD simulation.

An intrigued bias potential can be deduced considering equation (1.6.3). The Dirac delta function can be expressed by an its approximant:

$$\delta(x-a) = \lim_{\sigma\to 0} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-a)^2/(2\sigma^2)}$$

then substituting in equation (1.6.3) we have

$$Q(s) = \frac{1}{\sqrt{2\pi}} \lim_{t\to+\infty} \lim_{\sigma\to 0} \frac{1}{\sigma t} \int_0^t \exp\left(-\frac{(s(\vec{r}(\tau)) - s)^2}{2\sigma^2}\right) d\tau$$

the equation above suggest as a component of the history–dependent bias potential the use of a Gaussian function centered around the values assumed by the CVs at a time t. Then the bias potential component, extended to a multiple CVs, is chosen as follow

$$w(\vec{s}, t) = w \ \exp\left(-\sum_{i=0}^{n} \frac{(s_i - s_i(\vec{r}(t)))^2}{2\delta {s_i}^2}\right)$$

where w is the height and δs_i is the width of the deposited Gaussians.

For setting up a metadynamics simulation, then there are three parameters to carefully choose: the height w and the width δs_i of the deposited Gaussian and the stride of deposition τ. All these parameters affect the accuracy and the efficiency of the free energy profile reconstruction. Clearly if the Gaussians are big and large or placed quickly the FES will be explored at a fast pace but the reconstructed profile will be affected by large errors. Instead if they are small or placed infrequently the reconstruction will be accurate but will take a longer time. Moreover the time required to escape from a local minimum is determined by the number of Gaussians necessary to fill the well. This number is proportional to $(1/\delta s)^n$ [10], then the necessity to maintain the number n of CVs as small as possible or increase the Gaussian width. On the other hand the history–dependent potential can only reproduce features of the FES on a scale larger then $\sim \delta s$. Small criteria can be used to choose the δs and τ parameters; the former by monitoring the standard deviation of the CVs in an unbiased MD simulation and the latter by considering the relaxation time of the system after a Gaussian is deposited: clearly the bigger is the Gaussian the more time is needed by the system to well relax. However does not exist an universal and general recipe to choose these parameters, only some knowledge of the system and process under study can revel the right way; they can certainly fine tuned in successive steps.

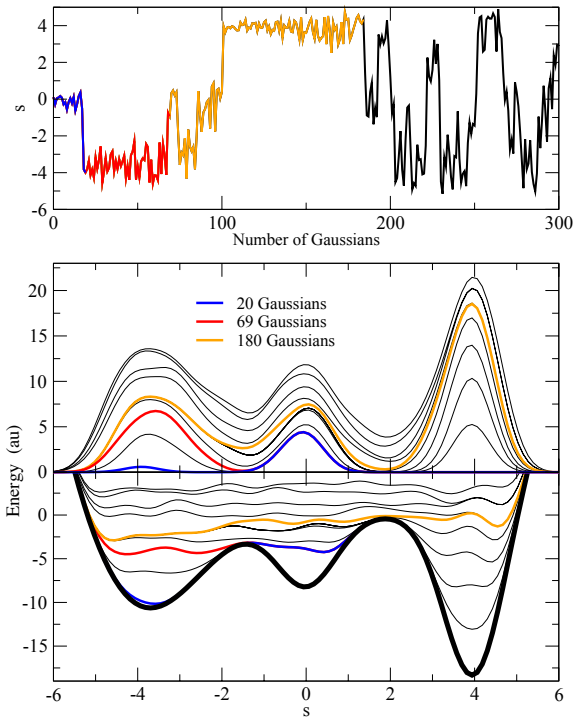


Figure 13: Example of a time evolution of a metadynamics simulation. Top: time evolution of the CV of a system evolved on the 3–minima FES represented by the black thick line in the lower panel. Middle: time evolution of w_M the history–dependent bias potential i.e. minus the FES. Blue line: w_M as when the first minimum is filled and the system escapes to the second minimum; red line: as when also the second minimum is filled; orange line: when the entire profile is filled and the dynamics becomes diffusive on the whole energy landscape. Lower panel: time evolution of the sum of w_M (same color scheme). Taken from [10].

In figure (13) an example of the time evolution of a metadynamics simulation is shown. We can see that as the deposited Gaussians increases the system is able to visit more regions of the phase space. Moreover as the simulation time increases, when the energy landscape (lower panel) become flat the system become diffusive (upper panel) on the whole energy profile. Thus we can stop the metadynamics simulation obtaining the right FES less then a constant as in equation (1.6.8).

Keeping in mind figure (13) a summary of the behavior of metadynamics and the validity of equation (1.6.8) can be qualitatively understood in the limit of slow deposition. This means that $w_M(\vec{s}, t)$ varies slowly and the probability to observe $\vec{s}$ is approximately proportional to the Boltzmann factor $e^{-\beta(G(\vec{s})-w_M(\vec{s},t))}$. If the function in the exponential has some local minimum due to $G(\vec{s})$, $\vec{s}$ will preferentially be localized in the neighborhood of this minimum and an increasing number of Gaussians have to be added until it is completely filled. When the minimum is filled the system reach the condition $G(\vec{s}) \sim w_M(\vec{s}, t)$ and the probability distribution would be approximately flat and we can say that the system become diffusive in the region explored by the simulation. Then, in this regions, the location of new Gaussians is no more affected by the difference between $G(\vec{s})$ and $w_M(\vec{s}, t)$ and they are deposited randomly over the flat free energy profile. Hence the FESs reconstructed after that point, as visible in figure (13), are affected by some corrugations, due to newly deposited Gaussians, of the order of w. From that point it says that the metadynamics has reached the *convergence* in the sampled region: the FES, as the Gaussians deposition increase, do not change its shape but it is only shifted downward along the energy axes.

CONVERGENCE As just explained before, the convergence of the metadynamics in a specific region of the CVs space is reached when the system become diffusive in this region, i.e. when the CVs can assume all possible values compatible to the sampled region. This is a crucial point for the metadynamics method in order to obtain the best possible estimator of the FES, in the sampled region. Clearly, in order to save computational time, if one want only to give the possibility to the system to escape from an energy minimum is not necessary to reach the convergence otherwise it is an important point. Unfortunately in most system is not trivial to identify precisely when the diffusive regime has reached. In most cases a practical method to assess the convergence of metadynamics simulations is based on monitoring the free energy difference between two reference point: when the difference is approximately flat then the convergence should be reached.

ERROR ESTIMATION After the proper convergence is reached the error on the FES estimator is clearly dependent on the chosen parameters. The best way to estimate the statistical error introduced by the metadynamics is to perform several statistically independent metadynamics runs. Clearly the arithmetic average of all the history–dependent potentials taken at the same time, after the convergence is reached, is still equal to the FES. The statistical error of the FES of a run can be considered as the standard deviation between the history–dependent potential $w_M(s,t)$ of that run and the averaged FES $G(\vec{s}) \sim -\overline{w_M(\vec{s},t)}$ and average it over the whole CVs space, as follow

$$\bar{\epsilon}^2 = \frac{1}{\Omega_s}\int_{\Omega_s} \overline{(w_M(\vec{s},t) - \overline{w_M(\vec{s},t)})^2}\, ds$$

where Ω_s is the whole CVs space. Laio *et al.* in [12], by performing extensive numerical simulations of a Langevin stochastic system, have derived an approximate expression for the error estimator in function of the system and metadynamics parameters, as follow

$$\bar{\epsilon}^2 \propto \frac{LT}{D}\frac{w}{\tau}\delta s$$

where L is the size of the simulation box, D is the system diffusion coefficient and T is the system temperature. Since δs is approximately fixed by the fluctuation of the CVs in a unbiased MD run and/or by the granularity to be achieved in the FES estimator, the error is dominated by the ration w/τ. Thus a fine tuning of the Gaussians height and deposition pace is often needed in order to minimize the statistical error. Despite this, is commonly accepted to follow a procedure for the FES and error estimators:

- Several statistically independent metadynamics simulations are performed;
- After the convergence is reached for each run, the estimator of the FES is the arithmetic average of the FES obtained from equation (1.6.1) for each simulations (each free energy profile appropriately normalized);

- The statistical error on the averaged FES is performed considering its standard deviation.

Alternatively one can perform a very long metadynamics run in which, for example, the system is diffusive for half simulation. Then one can chose as statistically independent history–dependent potentials some of them computed at different times in the diffusive region and follow the above procedure from the second point. Clearly one have to be careful about the decorrelation time between the chosen profiles, otherwise they are not statistically independent due the continuos nature of the metadynamics algorithm.

PERFORMANCE OPTIMIZATION In general the computational overhead of adding metadynamics to an MD simulation is usually not excessive, even if it depends on the implemented algorithms. However as the deposited Gaussians increases, each MD step a larger and larger number of exponential terms have to be computed and summed in order to calculate the derivative of the history–dependent potential, i.e. the forces due to the metadynamics. Thus if the Gaussians are big or frequently deposited or the system takes a lot of time to reach the convergence, the computational overhead scales as the number of the deposited Gaussians. This can clearly leads to a lose of performance as the simulation time increase. A simple solution is to implement a discrete mesh grid on the CVs space in which the history–dependent potential is spread and stored into. When a new Gaussian is added the potential is updated in the whole grid. While, at each MD step, in order to compute the derivative, the potential in a non–grid point is only estimated from an interpolation of some neighbor grid points. By this trick the computational overhead remains approximately constant as the simulation time increase.

1.6.3 Umbrella sampling and metadynamics remarks

Despite metadynamics is relatively recent while umbrella sampling is a well known and optimized method the former is steadily spread in the computational community against the latter. The most relevant reasons are its simplicity of implementation and the direct way to control efficiency and performance by changing the parameters of the Gaussians entering in the history–dependent potential. This gives to the metadynamics the possibility, with only one framework, to overcome different situations such as passing continuously from a fast and coarse exploration of the energy landscape to an accurate evaluation of the free energy profile, predict new stable configurations and structures, new reaction pathways, calculate free energy profiles and free energy differences.

In umbrella sampling the reconstruction of the free energy profile follow a predefined scheme designed for covering the chosen CVs space. In contract a well implemented metadynamics reconstruct efficiently the free energy profile. Starting from the current minimum and explore a larger and larger region of the accessible phase space dwelling on the low energy regions, which are statistically the most relevant, avoiding spend to much time in the irrelevant regions, the FES is recur-

sively reconstruct. Moreover which the advantage that each point in the CVs space is explored several times during the simulation. However a careful choice of the CVs is needed otherwise the history–dependent potential (and the reconstructed FES) can evolves in an unpredictable manner.

In a recent work by Davide Bochicchio *et al.* [2] both methods for the FES estimation included the error estimation, performance and accuracy are extensively compared with MD simulations involving the transfer of a hydrophobic oligomers from water phase to hydrophobic core of a lipid membrane. The authors consider both at atomistic and CG levels (MARTINI FF). The authors found that, if the CVs are properly chosen and the parameters for the metadynamics are carefully tuned, the energy profiles reconstructed are reasonable identical between umbrella sampling and metadynamics, but the latter yields the same accuracy in a less time consuming simulation.

2 | DUE

3 TRE

BIBLIOGRAPHY

[1] M.P. Allen and D.J. Tildesley. *Computer Simulation of Liquids*. Oxford Science Publ. Clarendon Press, 1989. ISBN: 9780198556459. URL: https://books.google.it/books?id=O32VXB9e5P4C.

[2] Davide Bochicchio et al. "Calculating the free energy of transfer of small solutes into a model lipid membrane: Comparison between metadynamics and umbrella sampling". In: *The Journal of Chemical Physics* 143.14, 144108 (2015). DOI: http://dx.doi.org/10.1063/1.4932159. URL: http://scitation.aip.org/content/aip/journal/jcp/143/14/10.1063/1.4932159.

[3] Giovanni Bussi, Davide Donadio, and Michele Parrinello. "Canonical sampling through velocity rescaling". In: *The Journal of Chemical Physics* 126.1, 014101 (2007). DOI: http://dx.doi.org/10.1063/1.2408420. URL: http://scitation.aip.org/content/aip/journal/jcp/126/1/10.1063/1.2408420.

[4] G. Andrés Cisneros et al. "Classical Electrostatics for Biomolecular Simulations". In: *Chemical Reviews* 114.1 (2014), pp. 779–814. DOI: http://dx.doi.org/10.1021/cr300461d. URL: http://dx.doi.org/10.1021/cr300461d.

[5] Tom Darden, Darrin York, and Lee Pedersen. "Particle mesh Ewald: An N · ln(N) method for Ewald sums in large systems". In: *The Journal of Chemical Physics* 98.12 (1993), pp. 10089–10092. DOI: http://dx.doi.org/10.1063/1.464397. URL: http://scitation.aip.org/content/aip/journal/jcp/98/12/10.1063/1.464397.

[6] Ulrich Essmann et al. "A smooth particle mesh Ewald method". In: *The Journal of Chemical Physics* 103.19 (1995), pp. 8577–8593. DOI: http://dx.doi.org/10.1063/1.470117. URL: http://scitation.aip.org/content/aip/journal/jcp/103/19/10.1063/1.470117.

[7] D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Computational science series. Elsevier Science, 2001. ISBN: 9780080519982. URL: https://books.google.it/books?id=5qTzldS9ROIC.

[8] Jochen S. Hub, Bert L. de Groot, and David van der Spoel. "g_wham-A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates". In: *Journal of Chemical Theory and Computation* 6.12 (2010), pp. 3713–3720. DOI: 10.1021/ct100494z. URL: http://dx.doi.org/10.1021/ct100494z.

[9] Johannes Kästner. "Umbrella sampling". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.6 (2011), pp. 932–942. ISSN: 1759-0884. DOI: 10.1002/wcms.66. URL: http://dx.doi.org/10.1002/wcms.66.

[10] Alessandro Laio and Francesco L. Gervasio. "Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science". In: *Reports on Progress in Physics* 71.12 (2008), p. 126601. URL: http://stacks.iop.org/0034-4885/71/i=12/a=126601.

[11] Alessandro Laio and Michele Parrinello. "Escaping free-energy minima". In: *Proceedings of the National Academy of Sciences* 99.20 (2002), pp. 12562–12566. DOI: 10.1073/pnas.202427399. URL: http://www.pnas.org/content/99/20/12562.abstract.

[12] Alessandro Laio et al. "Assessing the Accuracy of Metadynamics". In: *The Journal of Physical Chemistry B* 109.14 (2005), pp. 6714–6721. DOI: 10.1021/jp045424k. URL: http://dx.doi.org/10.1021/jp045424k.

[13] A. R. Leach. *Molecular Modelling: Principles and Applications*. Pearson Education. Prentice Hall, 2001. ISBN: 9780582382107. URL: https://books.google.it/books?id=kB7jsbV-uhkC.

[14] Cesar A. López et al. "Martini Coarse-Grained Force Field: Extension to Carbohydrates". In: *Journal of Chemical Theory and Computation* 5.12 (2009), pp. 3195–3210. DOI: 10.1021/ct900313w. URL: http://dx.doi.org/10.1021/ct900313w.

[15] S. J. Marrink et al. "The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations". In: *The Journal of Physical Chemistry B* 111.27 (2007), pp. 7812–7824. DOI: http://dx.doi.org/10.1021/jp071097f. URL: http://dx.doi.org/10.1021/jp071097f.

[16] Siewert J. Marrink and D. Peter Tieleman. "Perspective on the Martini model". In: *Chem. Soc. Rev.* 42 (16 2013), pp. 6801–6822. DOI: 10.1039/C3CS60093A. URL: http://dx.doi.org/10.1039/C3CS60093A.

[17] Luca Monticelli et al. "The MARTINI Coarse-Grained Force Field: Extension to Proteins". In: *Journal of Chemical Theory and Computation* 4.5 (2008), pp. 819–834. DOI: http://dx.doi.org/10.1021/ct700324x. URL: http://dx.doi.org/10.1021/ct700324x.

[18] M. Parrinello and A. Rahman. "Crystal Structure and Pair Potentials: A Molecular-Dynamics Study". In: *Phys. Rev. Lett.* 45 (14 Oct. 1980), pp. 1196–1199. DOI: 10.1103/PhysRevLett.45.1196. URL: http://link.aps.org/doi/10.1103/PhysRevLett.45.1196.

[19] M. Parrinello and A. Rahman. "Polymorphic transitions in single crystals: A new molecular dynamics method". In: *Journal of Applied Physics* 52.12 (1981), pp. 7182–7190. DOI: http://dx.doi.org/10.1063/1.328693. URL: http://scitation.aip.org/content/aip/journal/jap/52/12/10.1063/1.328693.

[20] Giulia Rossi et al. "Coarse-graining polymers with the MARTINI force-field: polystyrene as a benchmark case". In: *Soft Matter* 7 (2 2011), pp. 698–708. DOI: 10.1039/C0SM00481B. URL: http://dx.doi.org/10.1039/C0SM00481B.

[21] Benoît Roux. “The calculation of the potential of mean force using computer simulations”. In: *Computer Physics Communications* 91.1 (1995), pp. 275–282. ISSN: 0010-4655. DOI: http://dx.doi.org/10.1016/0010-4655(95)00053-I. URL: http://www.sciencedirect.com/science/article/pii/001046559500053I.

[22] M. Tuckerman. *Statistical Mechanics: Theory and Molecular Simulation*. Oxford Graduate Texts. OUP Oxford, 2010. ISBN: 9780191523465. URL: https://books.google.it/books?id=Lo3Jqc0pgrcC.

[23] L. Verlet. “Computer "Experiments" on Classical Fluids. II. Equilibrium Correlation Functions”. In: *Phys. Rev.* 165 (1 Jan. 1968), pp. 201–214. DOI: 10.1103/PhysRev.165.201. URL: http://link.aps.org/doi/10.1103/PhysRev.165.201.

[24] Semen O. Yesylevskyy et al. “Polarizable Water Model for the Coarse-Grained MARTINI Force Field”. In: *PLoS Comput Biol* 6.6 (June 2010), pp. 1–17. DOI: 10.1371/journal.pcbi.1000810. URL: http://dx.doi.org/10.1371/journal.pcbi.1000810.